

# Dictionary Learning Strategies for Compressed Fiber Sensing Using a Probabilistic Sparse Model

Christian Weiss and Abdelhak M. Zoubir

*Signal Processing Group, Institute of Communications and Graduate School of Computational Engineering,  
Technische Universität Darmstadt, Darmstadt, Germany*

October 24, 2016

## Abstract

We present a sparse estimation and dictionary learning framework for compressed fiber sensing based on a probabilistic hierarchical sparse model. To handle severe dictionary coherence, selective shrinkage is achieved using a Weibull prior, which can be related to non-convex optimization with  $\ell_p$ -norm constraints for  $0 < p < 1$ . In addition, we leverage the specific dictionary structure to promote collective shrinkage based on a local similarity model. This is incorporated in form of a kernel function in the joint prior density of the sparse coefficients, thereby establishing a Markov random field-relation. Approximate inference is accomplished using a hybrid technique that combines Hamilton Monte Carlo and Gibbs sampling. To estimate the dictionary parameter, we pursue two strategies, relying on either a deterministic or a probabilistic model for the dictionary parameter. In the first strategy, the parameter is estimated based on alternating estimation. In the second strategy, it is jointly estimated along with the sparse coefficients. The performance is evaluated in comparison to an existing method in various scenarios using simulations and experimental data.

## 1 Introduction

Fiber sensors are versatile devices with broad applicability [23, 38, 46, 63]. They are of high interest in smart structures to sense and react to the environment [42, 58]. For quasi-distributed sensing based on wavelength-division multiplexing (WDM), fiber Bragg grating (FBG) sensors are often employed due to their sensitivity to strain or temperature [23, 38]. An FBG describes a local variation of the refractive index and reflects light at a certain wavelength, called *Bragg wavelength*. Typically, a number of detuned FBGs is imprinted into the core of an optical fiber. Fiber interrogation is performed using broadband light sources or wavelength-tunable lasers. The latter feature higher local signal-to-noise ratios (SNRs) [46, 63]. However, in order to monitor time-varying perturbations, the laser has to sweep quickly through the tuning range. This requires high-speed analog-to-digital converters (ADCs) and produces large amounts of data.

*Compressed sensing* (CS) [5, 15, 27] can help to alleviate these problems by taking samples in form of projections into a low-dimensional subspace. The original signal can be reconstructed by exploiting the sparsity of the signal with respect to an adequate dictionary [11, 14]. This task strongly resembles the *sparse synthesis* problem with redundant dictionaries in [53, 57]. Besides greedy methods, such as Orthogonal Matching Pursuit (OMP) [49],  $\ell_1$ -minimization is a popular method to solve the sparse reconstruction problem [12, 24, 56]. It relies on the *restricted isometry property* (RIP), which essentially states that unique sparse solutions can be recovered by restricting the  $\ell_1$ -norm instead of the  $\ell_0$ -norm [13, 25]. Redundant dictionaries can yield highly sparse representations, that allow for estimating quantities at high resolution directly in the sparse domain [24, 41]. However, redundancy causes inter-column coherence and it is likely that the required RIP conditions are no longer fulfilled [12, 24, 52]. The  $\ell_p$ -norm, with  $0 < p < 1$ , offers a trade-off to avoid an NP-hard combinatorial problem imposed by the  $\ell_0$ -norm, while a unique solution might still be retrieved [17, 18].

Dictionaries can be classified as parametric or non-parametric. Non-parametric dictionaries are typically learned from training data and often used if no analytical model is available [53]. While they can yield sparser representations of certain data realizations [40], non-parametric dictionaries usually lack an interpretable structure and are inefficient in terms of storage [53]. Parametric dictionaries, in turn, rely on an analytical model for the observed signal. Their analytic form offers an efficient implementation and a means to obtain optimality proofs and error bounds [53]. They are also favorable in terms of scalability and storage-efficiency [4, 53]. *Translation-invariant* dictionaries represent an important sub-class of parametric dictionaries, that can be used to estimate the translation coefficients of localized signals [9, 29, 36]. Nonetheless, due to the complexity of natural signals, some model parameters might be unknown or contain uncertainty. Parametric *Dictionary Learning* (DL) addresses this problem with the aim of estimating these parameters from the measured data. Herein, statistical DL methods, such as maximum likelihood (ML) or maximum *a posteriori* (MAP) estimation, are commonly employed [53]. In order to solve the resulting optimization problem, *alternating estimation* (AE) is a frequently pursued sub-optimal paradigm, that iteratively optimizes a local objective function [2, 6, 43]. In a Bayesian setting, the Expectation Maximization (EM) algorithm is a popular variant of AE-based estimation [53].

A model for the sparse coefficients can be of deterministic or probabilistic nature. While the deterministic case is often assumed in sparse estimation [12, 24], a probabilistic model offers high flexibility to take model deviations and measurement errors into account. Moreover, a hierarchical structure can be used to incorporate additional uncertainty in prior assumptions. Sparsity can either be promoted by continuous distributions, resulting in *weakly sparse* models, or by discrete mixtures, leading to *strongly sparse* models [43]. A prominent example of discrete mixtures are *Spike & Slab* models [34]. They are based on binary activations and yield strongly sparse representations. Continuous sparse priors, such as a Gaussian or double-exponential (Laplace) prior, feature high excess kurtosis with heavy tails and a narrow peak around zero [43, 50]. Besides sparsity, additional knowledge of the signal, e.g. correlation, can be incorporated [26, 66].

For many practical models, evaluating the posterior distribution is not feasible and approximate methods, such as *Markov Chain Monte Carlo* (MCMC) or variational Bayes methods, have to be used to accomplish inference [8, 48, 54]. Variational methods use rather simple analytic functions to approximate the posterior distribution by factorization, which is favorable in terms of scalability and computational costs but leads to a deterministic approximation [8, 54]. MCMC methods attempt to sample the posterior distribution, where subsequent samples form a *Markov chain* [8]. The *Hamilton Monte Carlo* (HMC) method is a powerful technique, that is especially suitable for sampling high-dimensional spaces in the presence of correlation [47]. However, MCMC performance is generally limited by the available computation time, thereby relying on a stochastic approximation. Another application of MCMC is found in non-convex optimization, where Stochastic Gradient (SG) MCMC has gained popularity for large-scale Bayesian learning [19, 20, 22].

In the present work, we consider the problem of *Compressed Fiber Sensing* (CFS) with highly coherent translation-invariant dictionaries and imperfectly known parameters. For the sparse coefficients, a weakly sparse hierarchical model is considered. We also establish a relation between this model and non-convex optimization with  $\ell_p$ -norm constraints for  $0 < p < 1$ . In order to alleviate the problem of dictionary coherence, we leverage additional structure of the dictionary and achieve augmented sparsity by establishing a *Markov random field* (MRF) relation among the sparse coefficients. For dictionary learning, we pursue two different strategies: In the first strategy (**S1**), we consider a deterministic dictionary parameter, that is estimated using a Monte Carlo EM algorithm. In the second strategy (**S2**), a probabilistic hierarchical model for the dictionary parameter is considered, leading to a full Bayesian formulation and joint estimation of the sparse coefficients and the dictionary parameter. In both strategies, approximate inference is accomplished using a hybrid MCMC method based on Gibbs sampling and HMC. Finally, we use simulations and real data to compare the proposed methods to previous work in [62], where a deterministic model is considered for the sparse coefficients and the dictionary parameter. For the deterministic case, we derive the Cramér-Rao bound (CRB) to assess the performance gain achieved by a probabilistic model.

## 1.1 Contributions

- (I) We propose a probabilistic model for the sparse coefficients, where a Weibull prior is used to promote (weak) sparsity. Additional collective shrinkage is achieved by establishing an MRF-relation among the sparse coefficients based on a bivariate kernel function in the joint prior density. This helps to moderate the impact of severe dictionary coherence and can be used in general sparse synthesis problems with similar dictionary structure. We also establish a relation to non-convex optimization with constraints on the  $\ell_p$ -norm for  $0 < p < 1$ .
- (II) For dictionary learning, we investigate two conceptually different strategies, assuming either a deterministic (**S1**) or a stochastic (**S2**) dictionary parameter. In both strategies, the noise level can be jointly estimated along with the sparse coefficients. We further highlight advantages, disadvantages and limitations to offer support in choosing an adequate method for practical systems.
- (III) To accomplish inference in these models, we use a hybrid MCMC method, combining HMC and Gibbs sampling. We show its applicability and efficacy in the considered sampling problem for CFS.
- (IV) We use simulations to evaluate the performance of the proposed sparse estimation and DL methods for various scenarios of different CS sample sizes, SNRs and CS matrices. These results are compared to an existing method in [62], where the sparse coefficients and the dictionary parameter are assumed to be deterministic. In addition, we provide a real-data example to verify the practical applicability of **S1** and **S2**.
- (V) We derive the Cramér-Rao bound for jointly estimating the sparse coefficients and the dictionary parameter in the deterministic case. It is a valid bound for the competing method in [62], and serves to assess the achieved performance gain of our probabilistic approach.

## 2 Related Work

There exists little work addressing the combination of CS and DL for the application of WDM-based distributed fiber-optic sensing [60–62]. In [60], a model for the received sensor signal is presented, from which a redundant shift-invariant parametric dictionary is created. The works in [61, 62] focus on the aspect of CS and sparse estimation in the case of uncertain dictionary parameters. The authors use AE-based estimation to determine the dictionary parameters, where a pre-processing routine accounts for severe dictionary coherence. Unlike our approach, these works use a deterministic model for the sparse coefficients and dictionary parameters.

Weakly sparse models have been widely used in the literature. A comprehensive analysis of different hierarchical sparse prior models is provided in [44]. The general problem of choosing the prior in weakly sparse models for sparse regression is addressed in [50], where the authors describe various properties of different shrinkage priors and illuminate the selection problem from two perspectives: prior distributions and penalty functions. The work in [43] also investigates Bayesian methods with different sparse models in comparison to classical  $\ell_1$ -minimization. Seeger [54] found that the Laplace prior is able to shrink most components close to zero, while allowing for selected components to become sufficiently large. This effect, termed *selective shrinkage* in [35], is most noticeable for heavy-tailed priors, e.g. the Student’s  $t$ -prior [54] or the *horseshoe* prior in [16, 50]. Based on these findings, we select a sparsity prior that resembles a positive version of the horseshoe prior. Other works, that focus on the perspective of penalized regression, report higher sparsity levels by penalizing the  $\ell_p$ -norm with  $0 < p < 1$  instead of the  $\ell_1$ -norm [30]. The authors in [17] show that the RIP requirements for the dictionary can be relaxed in this case. It is also pointed out in [17, 18] that non-convex CS with  $\ell_p$ -norm penalization requires less measurements than standard CS, which is based on the  $\ell_1$ -norm. We rely on these results and show a relation between the considered sparsity prior and non-convex optimization with  $\ell_p$ -norm constraints.

There exist several approaches to exploit additional structure of the signal. One example is *block sparsity* [26]. A block sparse Bayesian learning framework is proposed in [66], pointing out how correlation can be exploited in regularization algorithms. Wakin *et al.* [59] introduce the concept of *joint sparsity* for signal recovery in distributed CS theory. In [41], temporal correlation across subsequent CS measurements is considered, while the authors in [21] use correlation to achieve smoothness. Another related concept is proposed in [3], where a truncated multivariate Ising MRF model is used to describe the correlation between adjacent pixels for image processing. Different from these works, we use the ideas of MRFs [45] and exploit correlation to achieve collective shrinkage among the sparse coefficients. A comparative analysis in [43] suggests that MCMC methods are powerful for inference in sparse models. In [47], the benefits of HMC and Gibbs sampling in hierarchical models are outlined. It is also shown, that HMC can be more effective than a Gibbs sampler for sampling high-dimensional spaces in the presence of correlation. According to these results, we consider a hybrid MCMC method that combines HMC and Gibbs sampling for inference in our hierarchical model, where the sparse coefficients are high-dimensional and correlated. For parametric DL, the Monte Carlo EM algorithm in **S1** represents one variant of the frequently applied AE-based estimation technique [39, 51]. Comparable to **S2** is the Bayesian framework for sparse estimation and DL in [31]. However, the authors use a Gaussian prior without correlation.

## 2.1 Outline

In Section 3, the signal model for CFS is introduced, and in Section 4, the CRB for joint estimation of the deterministic sparse coefficients and dictionary parameters is derived. Section 5 details the sparsity and local similarity model, while Section 6 describes the hybrid MCMC method for approximate inference in this model. The parametric DL strategies **S1** and **S2** are described in Section 7. Section 8 shows the working principle along with a performance analysis of the proposed and an existing method based on simulations and experimental data. A discussion of the results and findings is given in Section 9. Section 10 concludes this work.

## 3 Signal Model

In order to determine the quantity and nature of impairments at the FBGs in a WDM-based fiber sensor, the time delays of the reflections from the individual FBGs need to be estimated. We adopt the model in [60–62], where CS-based acquisition is employed to reduce the number of samples to be stored and processed. The CS measurements are described by

$$\mathbf{y} = \Phi \mathbf{A}(\theta) \mathbf{x} + \mathbf{n}, \quad (1)$$

where  $\Phi \in \mathbb{R}^{M \times L}$  is the CS sampling matrix and  $\mathbf{n} \in \mathbb{R}^M$  is a Gaussian noise component with independent and identically distributed (i.i.d.) entries,  $n_m \sim \mathcal{N}(0, \sigma_n^2)$ ,  $m = 1, \dots, M$ . The vector  $\mathbf{x} \in \mathbb{R}^N$  is sparse with  $K$  significant components, and  $\theta \in \mathbb{R}$  is a scalar dictionary parameter. The matrix  $\mathbf{A}(\theta)$  represents a redundant shift-invariant dictionary and its columns, called *atoms*, represent FBG reflections on a dense grid of delays. The indices of the  $K$  significant components in  $\mathbf{x}$  indicate the desired reflection delays. They are collected in the set  $\mathcal{S} = \{i_1, \dots, i_K\}$ . We can write the full data likelihood function for this model by

$$p(\mathbf{y} | \mathbf{x}, \theta) = (\sqrt{2\pi}\sigma_n)^{-M} \exp\left(-\frac{1}{2\sigma_n^2} \|\mathbf{y} - \Phi \mathbf{A}(\theta) \mathbf{x}\|_2^2\right). \quad (2)$$

The  $i$ -th dictionary atom,  $i = 1, \dots, N$ , is defined by [62]

$$[\mathbf{a}_i]_l(\theta) = r(lT_d - i\delta t, \theta), \quad l = 1, \dots, L, \quad (3)$$

where the generating function,  $r(lT_d - i\delta t, \theta)$ , describes the reflection from a single FBG, incrementally shifted by  $\delta t$  and sampled with a design sampling period,  $T_d$ . In order to specify the dictionary

parameter in CFS according to [62], we write

$$r(t, \theta) = \int_{-\infty}^{\infty} e^{j2\pi ft} H_{\text{LP}}(f, \theta) i_{\text{ph}}(f) df. \quad (4)$$

Herein,  $i_{\text{ph}}(f)$  is the received photocurrent in the frequency domain, and  $H_{\text{LP}}(f, \theta)$  is the transfer function of a lowpass filter, that models a limited *effective bandwidth* of the receiver circuitry. This bandwidth is described in terms of a positive dictionary parameter,  $\theta \in \mathbb{R}_+$ . As an auxiliary parameter, it accounts for different indistinguishable sources of uncertainty, that all contribute to the broadening in the temporal response of the FBG reflections. A detailed model for  $i_{\text{ph}}(f)$  is provided in [62].

## 4 The CRB for joint estimation of $(\mathbf{x}, \theta)$ in CFS

We derive the CRB for jointly estimating the deterministic parameters  $(\mathbf{x}, \theta)$ . This is a valid bound for the model considered in [62] and can be used to assess the relative performance gain achieved by the proposed probabilistic sparse model and DL strategies. Although the Bayesian CRB in [10] can be empirically determined, we found that this bound is very loose, due to the high information content in the considered sparsity prior. Therefore, and in regard of the comparative analysis with the deterministic case in [62], the non-Bayesian CRB is more useful in this case.

The constrained CRB for estimating  $\mathbf{x}$  with sparsity constraints has been derived in [7]. However, this derivation does not assume uncertainty in the dictionary. It is based on locally balanced sets and involves the projection of the *Fisher Information matrix* (FIM),  $\mathcal{I}(\mathbf{x})$ , onto a low-dimensional subspace spanned by the so-called *feasible directions*. Any estimator,  $\hat{\mathbf{x}}$ , for which the constrained CRB is a valid lower bound, must be unbiased with respect to these directions. The projection matrix can be created from the unit vectors corresponding to the non-zero coefficients in  $\mathbf{x}$ , that is  $\mathbf{U} = [\mathbf{e}_{i_1}, \dots, \mathbf{e}_{i_K}]$  with  $i_k \in \mathcal{S}$ ,  $k = 1, \dots, K$ . For a Gaussian likelihood as in (2), the FIM can be derived from the expected value of the Hessian matrix of the log-likelihood function, i.e. [7, 37]

$$\mathcal{I}(\mathbf{x}) = -\mathbb{E}_{\mathbf{y}} \nabla_{\mathbf{x}}^2 \log p(\mathbf{y} | \mathbf{x}, \theta) = \frac{1}{\sigma_n^2} \mathbf{B}^\top \mathbf{B}, \quad (5)$$

with  $\mathbf{B} = \Phi \mathbf{A}$ . Further, we define the reduced FIM by  $\mathcal{I}_K := \mathbf{U}^\top \mathcal{I}(\mathbf{x}) \mathbf{U}$ . Then, given that  $\mathbf{x}$  is exactly  $K$ -sparse, the constrained CRB for a *known* dictionary becomes [7]

$$\text{Cov}(\hat{\mathbf{x}}) \succeq \mathbf{U} \mathcal{I}_K^{-1} \mathbf{U}^\top, \quad \|\mathbf{x}\|_0 = K. \quad (6)$$

Based on these results, we derive the CRB for the joint parameters  $\boldsymbol{\gamma} = (\mathbf{x}, \theta)$ . First, we derive the Fisher information for  $\theta$ , given that  $\mathbf{x}$  is known. It is given by

$$\begin{aligned} \mathcal{I}(\theta) &= -\mathbb{E}_{\mathbf{y}} \frac{\partial^2}{\partial \theta^2} \log p(\mathbf{y} | \mathbf{x}, \theta) \\ &= \mathbb{E}_{\mathbf{y}} \frac{\partial^2}{\partial \theta^2} \frac{1}{2\sigma_n^2} (\mathbf{y} - \Phi \mathbf{A}(\theta) \mathbf{x})^\top (\mathbf{y} - \Phi \mathbf{A}(\theta) \mathbf{x}) \\ &= \frac{1}{\sigma_n^2} \mathbf{x}^\top \mathbf{A}'(\theta)^\top \Phi^\top \Phi \mathbf{A}'(\theta) \mathbf{x}. \end{aligned} \quad (7)$$

Herein,  $\mathbf{A}'(\theta)$  denotes the (element-wise) derivative of  $\mathbf{A}(\theta)$  with respect to  $\theta$ . Next, we have to take into account that  $\mathbf{x}$  and  $\theta$  share some mutual information. Therefore, we define the combined FIM:

$$\mathcal{I}(\boldsymbol{\gamma}) = \begin{pmatrix} \mathcal{I}(\mathbf{x}) & -\mathbb{E}_{\mathbf{y}} \mathbf{u} \\ -\mathbb{E}_{\mathbf{y}} \mathbf{u}^\top & \mathcal{I}(\theta) \end{pmatrix}, \quad (8)$$

where  $\mathbf{u} = [u_1, \dots, u_N]^\top$  and  $u_i = \frac{\partial}{\partial x_i} \frac{\partial}{\partial \theta} \log p(\mathbf{y} | \mathbf{x}, \theta)$ ,  $i = 1, \dots, N$ . Since the partial derivatives can be interchanged, the off-diagonal elements are identical. In order to complete the definition of  $\mathcal{I}(\boldsymbol{\gamma})$ ,

we determine

$$\begin{aligned}
-\mathbb{E}_{\mathbf{y}} u_i &= -\mathbb{E}_{\mathbf{y}} \frac{\partial^2}{\partial x_i \partial \theta} \log p(\mathbf{y} | \mathbf{x}, \theta) \\
&= \frac{1}{\sigma_n^2} \mathbf{x}^\top \mathbf{A}'(\theta)^\top \mathbf{\Phi}^\top \mathbf{\Phi} \mathbf{a}_i(\theta).
\end{aligned} \tag{9}$$

The reduced FIM is obtained by appending the set of feasible directions, such that the coordinate  $\theta$  is included, i.e.  $\tilde{\mathbf{U}} = [\mathbf{U}, \mathbf{e}_{N+1}]$ . Hence,  $\mathcal{I}_{K+1} = \tilde{\mathbf{U}}^\top \mathcal{I}(\gamma) \tilde{\mathbf{U}}$ . To obtain the inverse, we apply twice the matrix inversion lemma [32]

$$\mathcal{I}_{K+1}^{-1} = \begin{pmatrix} \left( \mathcal{I}_K - \frac{\mathbf{v}\mathbf{v}^\top}{\mathcal{I}(\theta)} \right)^{-1} & -\frac{1}{\check{b}} \mathcal{I}_K^{-1} \mathbf{v} \\ -\frac{1}{\check{b}} \mathbf{v}^\top \mathcal{I}_K^{-1} & \frac{1}{\check{b}} \end{pmatrix}, \tag{10}$$

where  $\check{b} = \mathcal{I}(\theta) - \mathbf{v}^\top \mathcal{I}_K^{-1} \mathbf{v}$ , and

$$\left( \mathcal{I}_K - \frac{\mathbf{v}\mathbf{v}^\top}{\mathcal{I}(\theta)} \right)^{-1} = \mathcal{I}_K^{-1} + \frac{1}{\check{b}} \mathcal{I}_K^{-1} \mathbf{v} \mathbf{v}^\top \mathcal{I}_K^{-1}. \tag{11}$$

The constrained CRB for the joint parameters in  $\gamma$  becomes

$$\text{Cov}(\gamma) \succeq \tilde{\mathbf{U}} \mathcal{I}_{K+1}^{-1} \tilde{\mathbf{U}}^\top, \quad \|\mathbf{x}\|_0 = K. \tag{12}$$

Finally, a lower bound for the *mean squared error (MSE)* in the joint setting is obtained by updating the individual estimation errors to account for the information shared between  $\mathbf{x}$  and  $\theta$ :

$$\text{MSE}(\hat{\mathbf{x}}) \geq (\text{Tr } \mathcal{I}_K^{-1}) + \frac{1}{\check{b}} \mathbf{v}^\top \mathcal{I}_K^{-1} \mathcal{I}_K^{-1} \mathbf{v} \tag{13}$$

$$\text{MSE}(\hat{\theta}) \geq \frac{1}{\check{b}} = \mathcal{I}(\theta)^{-1} + \frac{\mathbf{v}^\top \mathcal{I}_K \mathbf{v}}{\mathcal{I}(\theta) (\mathcal{I}(\theta) - \mathbf{v}^\top \mathcal{I}_K \mathbf{v})}. \tag{14}$$

## 5 Probabilistic sparse model

Regarding the model in (1), the data can be explained in different ways. On the one hand, many non-zero components in  $\mathbf{x}$  and a large bandwidth,  $\theta$ , result in many narrow temporal peaks that can yield a good approximation of the observed reflections. On the other hand, it is known that the sensing fiber contains  $K$  FBGs, so we expect exactly  $K$  reflections. Therefore, a more useful explanation is given by  $K$  significant elements in  $\mathbf{x}$  with a smaller value of  $\theta$ , such that  $\mathcal{S}$  correctly indicates the reflection delays. Nevertheless, even for a suitable value of  $\theta$ , the signal  $\mathbf{x}$  is usually not exactly sparse but contains many small elements close to zero, e.g. due to measurement noise. In a strongly sparse model, these contributions are not taken into account, which impacts the positions of non-zero elements in  $\mathbf{x}$ . Hence, it may lead to incorrectly estimated reflection delays. This motivates a weakly sparse model, where the  $K$  most significant components indicate the reflection delays. When  $\mathbf{x}$  and  $\theta$  are both unknown, the reflections delays can only be estimated when prior information of sparsity is incorporated, since  $\theta$  depends on  $\mathbf{x}$  and vice versa. Severe dictionary coherence aggravates this problem and results in several non-zero components with moderate amplitudes around the true significant elements. The coherence level is even stronger when the dimensionality of the acquired data is further reduced by CS. Thus, an adequate sparse model for  $\mathbf{x}$  must compensate for this effect. Classic  $\ell_1$ -minimization can be interpreted as an MAP estimation problem, where  $\mathbf{x}$  has i.i.d. entries with Laplace priors [43]. However, the required performance guarantees for  $\ell_1$ -minimization, essentially the RIP [11, 14], are no longer fulfilled in the case of strong dictionary coherence. According to [17, 18], the RIP conditions can be relaxed for  $\ell_p$ -minimization, when  $0 < p < 1$ . Therefore, we use a prior with stronger selective shrinkage effect, that can be related to constraints on the  $\ell_p$ -norm in non-convex optimization. Yet, specific characteristics of the signal have to be considered. The measured reflection

signal is proportional to the optical power, and the dictionary atoms essentially model the optical power reflected from the individual FBGs. Thus, the prior must also account for the non-negativity of the data. Due to these restrictions, we choose a Weibull prior that resembles a positive version of the horseshoe prior in [50] and induces the required selective shrinkage effect:

$$x_i \sim p(x_i) = \mathcal{W}(x_i | \lambda_w, k_w), \quad x_i > 0, \quad i = 1, \dots, N, \quad (15)$$

where  $\lambda_w, k_w$  are the scale and shape parameters, respectively. Then, the joint prior density of  $\mathbf{x}$  is given by

$$p(\mathbf{x} | k_w, \lambda_w) = \frac{k_w}{\lambda_w^{k_w}} \prod_{i=1}^N x_i^{k_w-1} \exp\left(-\lambda_w^{-k_w} \sum_{i=1}^N x_i^{k_w}\right). \quad (16)$$

Fig. 1 (top left) shows qualitatively the shape of the considered prior in the bivariate case.

Based on (16) and (2), we can relate the problem to constrained ML estimation. First, let us consider an interpretation in terms of MAP estimation as in [44], by calculating  $\arg \max_{\mathbf{x}} \log p(\mathbf{y} | \mathbf{x}, \theta) p(\mathbf{x} | k_w, \lambda_w)$  or, equivalently,

$$\begin{aligned} \arg \min_{\mathbf{x}} -\log p(\mathbf{y} | \mathbf{x}, \theta) p(\mathbf{x} | k_w, \lambda_w) = \\ \arg \min_{\mathbf{x}} \|\mathbf{y} - \Phi \mathbf{A}(\theta) \mathbf{x}\|_2^2 + \mu_1 \sum_{i=1}^N \log(x_i) + \mu_2 \sum_{i=1}^N x_i^{k_w}, \end{aligned} \quad (17)$$

where  $\mu_1 = (1 - k_w)$  and  $\mu_2 = \lambda_w^{-k_w}$  with  $0 < k_w < 1$  and  $\mu_1, \mu_2 > 0$ . In order to formulate a related constrained ML problem, let us define two functions,

$$g_1 = \sum_{i=1}^N x_i^{k_w} - \lambda_1^{k_w} \quad \text{and} \quad g_2 = \sum_{i=1}^N \log(x_i) - \lambda_2, \quad (18)$$

where  $\lambda_1, \lambda_2 \in \mathbb{R}_+$  are related to the coefficients  $\mu_1, \mu_2$ , respectively. The functions in (18) can represent inequality constraints of the form  $g_1 \leq 0$  and  $g_2 \leq 0$ , that account for the impact of the prior by restricting the search space. Hence, a constrained version of the ML problem can be formulated by

$$\arg \min_{\mathbf{x} > \mathbf{0}} \|\mathbf{y} - \Phi \mathbf{A}(\theta) \mathbf{x}\|_2^2 \quad (19)$$

$$\text{s.t.} \quad \|\mathbf{x}\|_{k_w} \leq \lambda_1 \quad (20)$$

$$\text{and} \quad \sum_{i=1}^N \log(x_i) \leq \lambda_2. \quad (21)$$

In this non-convex problem,  $\|\mathbf{x}\|_p = (\sum_{i=1}^N |x_i|^p)^{1/p}$  denotes the  $\ell_p$ -norm with  $p = k_w < 1$ . The hyper-parameters  $\lambda_1, \lambda_2$  control the shrinkage effects. Fig. 1 (top right) depicts the search space restricted by the constraints (20)-(21). The borders are shown for a fixed value of  $\lambda_1$  and  $\lambda_2$  in the bivariate case.

## 5.1 Local covariance model for augmented sparsity

In analogy to the concept of block sparsity [26], we can use the specific sparse structure of the signal with respect to the shift-invariant dictionary for CFS to exploit sparsity among groups of variables. The signal contains only  $K$  reflections that arrive at temporally separated delays, indicated by the significant components in  $\mathbf{x}$ . Therefore, we can assume that a significant coefficient is always surrounded by larger groups of non-significant coefficients and any two significant components are always well separated. Also, it is likely that the amplitudes of adjacent non-significant coefficients are similarly close to zero. Borrowing from the ideas of MRFs [45], such local similarity can be modeled by a prior on the differential coefficients,  $\Delta \mathbf{x}$ , where  $\Delta x_i = x_{i+1} - x_i$ ,  $i = 1, \dots, N-1$ . It restricts the variation of adjacent amplitudes and establishes a MRF relation between neighboring coefficients in  $\mathbf{x}$ . Then, non-significant coefficients with larger amplitudes are pulled down to match the majority with almost-zero amplitudes, which promotes additional *collective* shrinkage. However, if a significant coefficient

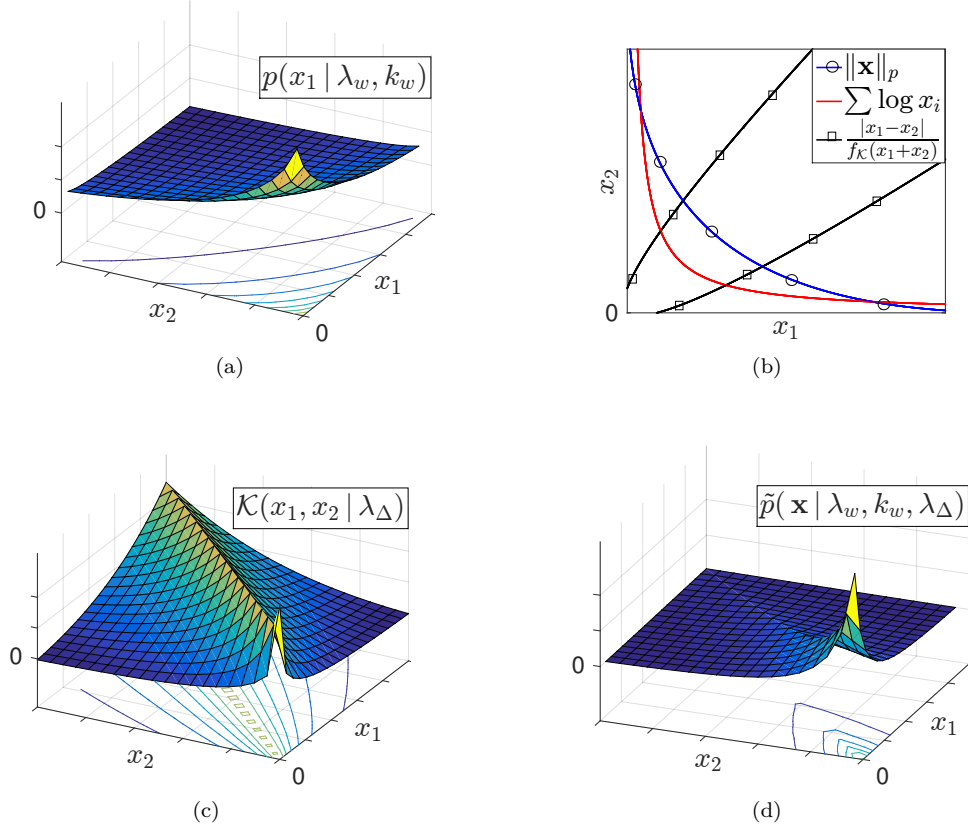


Figure 1: Bivariate examples. Top left: Weibull prior  $p(x_1 | \lambda_w, k_w)$ . Top right: search space of the constrained non-convex optimization problem. Bottom left: local similarity function  $\mathcal{K}(x_1, x_2 | \lambda_\Delta)$ . Bottom right: modified joint density  $\tilde{p}(\mathbf{x} | \lambda_w, k_w, \lambda_\Delta)$  with  $\mathbf{x} = [x_1, x_2]^\top$  and  $f_\kappa(x_1, x_2) = \sqrt{x_1 + x_2}$ .

follows a non-significant one (or vice versa), the model should allow for larger changes. Therefore, the differential variation must be locally specified, dependent on the respective amplitudes, in order to avoid undesired shrinkage or equalization. To this end, we define a kernel function for all adjacent pairs of sparse coefficients, i.e.  $\forall i = 1, \dots, N-1$ , with hyperparameter  $\lambda_\Delta$ :

$$\mathcal{K}(x_i, x_{i+1} | \lambda_\Delta) = \exp\left(-\lambda_\Delta \frac{|x_{i+1} - x_i|}{f_\kappa(x_i, x_{i+1})}\right). \quad (22)$$

The bivariate function  $f_\kappa$  controls the similarity level between adjacent coefficients. Within the scope of this work we consider cases, where this function takes the form  $f_\kappa(x_i, x_{i+1}) = (x_i + x_{i+1})^r / N_x$ ,  $i = 1, \dots, N-1$ , with positive constants  $r \leq 1$ ,  $N_x < \infty$ . They can be incorporated in  $p(\mathbf{x} | k_w, \lambda_w)$  to yield a modified joint prior density,

$$\begin{aligned} \tilde{p}(\mathbf{x} | k_w, \lambda_w, \lambda_\Delta) &= \frac{1}{Z_\kappa} \mathcal{W}(x_N | k_w, \lambda_w) \\ &\times \prod_{i=1}^{N-1} \mathcal{K}(x_i, x_{i+1} | \lambda_\Delta) \mathcal{W}(x_i | k_w, \lambda_w), \end{aligned} \quad (23)$$

with normalization constant  $Z_\kappa$ . For any  $\alpha, \beta \in \mathbb{R}_+$ , it holds that  $0 < \mathcal{K}(\alpha, \beta | \lambda_\Delta) = \mathcal{K}(\beta, \alpha | \lambda_\Delta) \leq 1$  and

$$\tilde{p}(\mathbf{x} | k_w, \lambda_w, \lambda_\Delta)|_{Z_\kappa=1} \leq p(\mathbf{x} | k_w, \lambda_w) \quad (24)$$



is bounded. Hence, there exists a positive constant  $Z_K < \infty$  that normalizes (23) to make  $\tilde{p}(\mathbf{x} | k_w, \lambda_w, \lambda_\Delta)$  a proper density. Fig. 1 (bottom) visualizes the function  $\mathcal{K}(x_i, x_{i+1} | \lambda_\Delta)$  and its impact on the original prior in the bivariate case.

In the view of constraint ML estimation, the modified prior density in (23) can be related to the optimization problem in (19)-(21) by imposing additional constraints

$$\frac{|x_{i+1} - x_i|}{f_K(x_i, x_{i+1})} \leq \mu_i, \quad i = 1, \dots, N-1. \quad (25)$$

Fig. 1 (top right) depicts a bivariate example. In order to show the MRF relation between the coefficients, we calculate the conditional densities  $\forall x_i, i = 1, \dots, N$ . To this end, we conveniently define  $p(x_i | \mathbf{x}_{\setminus i}) = p(x_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N)$  and get

$$\begin{aligned} \tilde{p}(x_i | \mathbf{x}_{\setminus i}, k_w, \lambda_w, \lambda_\Delta) &= \tilde{p}(x_i | x_{i-1}, x_{i+1}, k_w, \lambda_w, \lambda_\Delta) \\ &\propto \mathcal{W}(x_i | k_w, \lambda_w) \mathcal{K}(x_{i-1}, x_i | \lambda_\Delta) \mathcal{K}(x_i, x_{i+1} | \lambda_\Delta), \end{aligned} \quad (26)$$

$$\tilde{p}(x_1 | \mathbf{x}_{\setminus 1}, k_w, \lambda_w, \lambda_\Delta) \propto \mathcal{W}(x_1 | k_w, \lambda_w) \mathcal{K}(x_1, x_2 | \lambda_\Delta), \quad (27)$$

$$\tilde{p}(x_N | \mathbf{x}_{\setminus N}, k_w, \lambda_w, \lambda_\Delta) \propto \mathcal{W}(x_N | k_w, \lambda_w) \mathcal{K}(x_{N-1}, x_N | \lambda_\Delta), \quad (28)$$

where dependencies appear only between directly adjacent coefficients.

In order to account for deviations from prior assumptions, we consider randomization of the hyperparameters and assign conjugate inverse Gamma priors to the scale parameters  $\lambda_w$  and  $\lambda_\Delta$ . Finally, given  $\lambda_w$  and a normalization constant  $Z_{k_w}$ , the shape parameter,  $k_w > 0$ , is assigned the conjugate prior distribution according to [28]:

$$p(k_w | a', b', (d')^{k_w}, \lambda_w) = \frac{k_w^{a'}}{Z_{k_w}} \exp\left(-b' k_w - \frac{(d')^{k_w}}{\lambda_w}\right), \quad (29)$$

Fig. 2 shows a factor graph for the complete sparsity model with randomized hyperparameters.

## 6 Approximate Inference: Hybrid MCMC

In order to accomplish inference in the sparse model, we apply a hybrid MCMC technique, i.e. HMC within Gibbs sampling. The reasons for using HMC are twofold: Firstly, it only requires an analytic expression for the posterior density to be sampled. Secondly, it is efficient in sampling high-dimensional spaces in the presence of correlation. However, as pointed out in [47], it can be more efficient to sample the hyperparameters separately, as their posterior distributions are often highly peaked and require a small step size in the HMC algorithm, which limits the general performance. Therefore, we employ an outer Gibbs sampler for approximate inference of the latent variables. In each iteration,  $\tilde{p}(\mathbf{x} | \lambda_w, k_w, \lambda_\Delta)$  is sampled using HMC, while all other variables are fixed. Since we are also interested in estimating the noise variance,  $\sigma_n^2$ , it is assigned an inverse Gamma (Inv- $\Gamma$ ) prior and sampled along with the other variables. The resulting model is summarized below:

$$\begin{aligned} \mathbf{x} | k_w, \lambda_w, \lambda_\Delta &\sim \tilde{p}(\mathbf{x} | k_w, \lambda_w, \lambda_\Delta) && \text{in (23),} \\ \lambda_w &\sim \text{Inv-}\Gamma(\lambda_w | a, b), \\ k_w | \lambda_w &\sim p(k_w | a', b', (d')^{k_w}, \lambda_w) && \text{in (29),} \\ \lambda_\Delta &\sim \text{Inv-}\Gamma(\lambda_\Delta | a'', b'') \\ \sigma_n^2 &\sim \text{Inv-}\Gamma(\lambda_\Delta | a_\sigma, b_\sigma). \end{aligned} \quad (30)$$

We also define  $\zeta \in \mathcal{C} = \{k_w, \lambda_w, \lambda_\Delta, \sigma_n^2\}$  as a representative variable with corresponding positive, real-valued parameters  $a_\zeta \in \{a, a'', a_\sigma\}$  and  $b_\zeta \in \{a, a'', a_\sigma\}$ , that belong to the respective density functions. Further, the set  $\mathcal{C}_{\setminus \zeta}$  denotes the set  $\mathcal{C}$  without the respective variable  $\zeta$ . Fig. 3 shows a graphical model that helps to visualize the dependencies in this model. Herein,  $\theta$  and  $\Xi$  are only

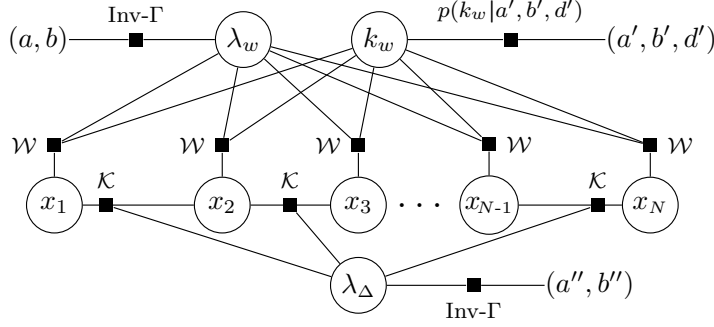


Figure 2: Factor graph of the complete sparse model with local similarity.

valid for strategy **S2**, which is discussed in Section 7. For the particular model in (30), we assume that the variables  $\mathbf{x}$ ,  $\sigma_n^2$  and  $\theta$  are mutually independent. Gibbs sampling requires the full conditional distributions for each parameter of interest. Based on these assumptions, we obtain the relation

$$p(\zeta | \mathbf{y}, \mathbf{x}, \mathcal{C}_{\setminus \zeta}) \propto p(\mathbf{y} | \mathbf{x}, \mathcal{C}) p(\zeta | \mathbf{x}, \mathcal{C}_{\setminus \zeta}) \quad (31)$$

$$\propto p(\mathbf{y} | \mathbf{x}, \mathcal{C}) p(\zeta | \mathcal{C}_{\setminus \zeta}) \tilde{p}(\mathbf{x} | \mathcal{C}). \quad (32)$$

Since the prior distributions are all conjugate to the Gaussian likelihood function in (2), a simple calculation yields the posterior distributions of the parameters involved in the Gibbs sampling procedure. For  $\zeta \in \mathcal{C}_{\setminus k_w}$ , we obtain

$$\zeta | \mathbf{y}, \mathbf{x}, \mathcal{C}_{\setminus \zeta} \sim \text{Inv-}\Gamma(\zeta | a_{\zeta} + \frac{M}{2}, b_{\zeta} + \frac{1}{2}) \tilde{p}(\mathbf{x} | \mathcal{C}), \quad (33)$$

and for  $k_w$ , we obtain

$$k_w | \mathbf{y}, \mathbf{x}, \mathcal{C}_{\setminus k_w} \sim p(k_w | \tilde{a}', \tilde{b}', \tilde{c}') \tilde{p}(\mathbf{x} | \mathcal{C}), \quad (34)$$

with parameters  $\tilde{a}' = a' + N$ ,  $\tilde{b}' = b' + \sum_{i=1}^N \log(x_i)$ , and  $\tilde{c}' = (d')^{k_w} + \sum_{i=1}^N x_i^{k_w}$ . Samples of the posterior variables can be obtained using Metropolis Hastings [8] or HMC.

The sparse coefficients are sampled using HMC. We briefly describe the idea of this method according to [47], adapted to our model for  $\mathbf{x}$ :

Within the framework of HMC, the sampling process is described in terms of *Hamilton dynamics*, a concept known from classical physics. It is used to describe the trajectory of a physical system in phase space, based on its potential and kinetic energy. HMC assigns to every sparse coefficient,  $x_i$ , an associated momentum variable,  $\xi_i$ ,  $i = 1, \dots, N$ , that is responsible for the sampling dynamics. The posterior density to be sampled is related to the potential energy, given by [47]

$$U(\mathbf{x} | \mathbf{y}, \mathcal{C}) = -\log \tilde{p}(\mathbf{x} | \mathbf{y}, \mathcal{C}) - \log(Z_u), \quad (35)$$

where  $Z_u$  is a suitable normalization constant. Since  $\mathbf{y}$  and  $\mathcal{C}$  are fixed, we may drop them and write  $U(\mathbf{x})$  instead. The kinetic energy,  $K(\boldsymbol{\xi})$ , depends only on the auxiliary variables  $\boldsymbol{\xi} = [\xi_1, \dots, \xi_N]$ . A standard choice for  $K(\boldsymbol{\xi})$  corresponds to independent particles in free space with mass  $m_i$ , i.e.  $K(\boldsymbol{\xi}) = \sum_{i=1}^N \xi_i^2 / 2m_i$ . The dynamics of the sampling process are governed by the *Hamiltonian function*, which is given by  $\mathcal{H}(\mathbf{x}, \boldsymbol{\xi}) = U(\mathbf{x}) + K(\boldsymbol{\xi})$  and represents the total system energy. The joint density of  $(\mathbf{x}, \boldsymbol{\xi})$  is defined by [47]

$$p(\mathbf{x}, \boldsymbol{\xi}) = \frac{1}{Z_c} e^{-\frac{\mathcal{H}(\mathbf{x}, \boldsymbol{\xi})}{T_{\text{sys}}}} = \tilde{p}(\mathbf{x} | \mathbf{y}, \mathcal{C}) \prod_{i=1}^N \mathcal{N}(\xi_i | 0, m_i). \quad (36)$$

Herein,  $T_{\text{sys}}$  is called the *system temperature* and  $Z_c$  is a normalization constant. The last equation is obtained by setting  $T_{\text{sys}} = 1$  and  $Z_u = Z_c$ , while the Gaussian density arises from the special

choice of the kinetic energy term. In HMC, a proposal for a new sample is obtained by the final points  $(x_i^*, \xi_i^*)$  of a trajectory described by Hamilton's equations of motion. They are calculated  $\forall (x_i, \xi_i), i=1, \dots, N$ , [47]:

$$\frac{dx_i}{dt} = \frac{\xi_i}{m_i}, \quad \frac{d\xi_i}{dt} = -\frac{\frac{\partial}{\partial x_i} \tilde{p}(\mathbf{x} | \mathbf{y}, \mathcal{C})}{\tilde{p}(\mathbf{x} | \mathbf{y}, \mathcal{C})}. \quad (37)$$

A Metropolis update decides, whether a proposed sample is accepted or rejected, with acceptance probability [47]

$$P(\text{accept}) = \min(1, \exp(-\mathcal{H}(x_i^*, \xi_i^*) + \mathcal{H}(x_i, \xi_i))) . \quad (38)$$

## 7 Parametric DL strategies for CFS

In this section, we present two strategies for parametric dictionary learning in CFS. In the first strategy (**S1**), we follow the ideas of hybrid Bayesian inference [64, 65] and AM-based DL [6], where  $\theta$  is a deterministic parameter, that is estimated using the Monte Carlo EM algorithm in [8]. In the second strategy (**S2**), we pursue a full Bayesian approach and consider a probabilistic model for  $\theta$ . Herein, approximate inference is accomplished by extending the Gibbs sampler in Section 6 to jointly estimate  $(\mathbf{x}, \theta, \sigma_n^2)$ . Fig. 3 depicts the dependency graph for both strategies, where  $\theta, \Xi$  belong exclusively to **S2**.

As pointed out in [64, 65], hybrid and full Bayesian strategies have their individual advantages in certain situations. For small sample sizes, Bayesian methods can be superior if good prior knowledge is available [64]. Nonetheless, they are often computationally more complex and insufficient prior information can lead to a small-sample bias, even if a non-informative prior is used [64]. In CFS, the sample size is small and only vague prior knowledge of  $\theta$  is available. Therefore, we investigate the performance of both DL strategies based on our probabilistic sparse model. The computational complexity of both strategies is comparable. It is dominated by HMC, i.e. by sampling the high-dimensional vector  $\mathbf{x}$  in each iteration of the Gibbs sampler. Regarding  $\theta$ , the following prior knowledge is assumed: In **S1**, we roughly restrict the range of values that  $\theta$  can take, while in **S2**, we define a non-informative prior over the same range. Recall that  $\theta$  effectively describes the filter characteristics of the lowpass filter  $H_{\text{LP}}(\omega)$ . To create the dictionary for a certain value of  $\theta$  using (3), the inverse Fourier transform in (4) has to be evaluated for each atom. Thus, the dictionary is not a simple function of  $\theta$  and we restrict ourselves to a discrete set of parameters, with lower and upper bound,  $\theta_{\min}$  and  $\theta_{\max}$ , respectively. Since the bandwidth should be positive and bounded, we have  $0 < \theta_{\min}$  and  $\theta_{\max} < \infty$ . Then, the set  $\Theta$  contains the discrete values  $\theta_r, r = 1, \dots, R_\Theta$ ,

### 7.1 Hybrid DL: iterative estimation of $\theta$ and $(\mathbf{x}, \mathcal{C})$ (**S1**)

The dictionary parameters in the CFS problem can be iteratively estimated using a Monte Carlo EM algorithm. First, an initial value,  $\theta^{(0)}$ , has to be chosen. In subsequent iterations with indices  $d = 1, \dots, d_{\max}$ , we obtain joint samples  $\{\mathbf{x}_l, \mathcal{C}_l\}^{(d)}, l = 1, \dots, L_{\text{MC}}$ , by Gibbs sampling and HMC according to Section 6. Then, we determine the posterior expectation of  $\zeta \in \mathcal{C}$ , using the previous estimate  $\hat{\theta}^{(d-1)}$ :

$$\hat{\zeta}^{(d)} = \int_{\text{dom}(\zeta)} \zeta p(\zeta | \mathbf{y}, \hat{\theta}^{(d-1)}) d\zeta \quad (39)$$

$$\approx \frac{1}{L_{\text{MC}}} \sum_{l=1}^{L_{\text{MC}}} \zeta_l^{(d)} p(\zeta_l^{(d)} | \mathbf{y}, \hat{\theta}^{(d-1)}), \quad (40)$$

where  $\text{dom}(\zeta)$  is the domain of  $\zeta$ . The current estimates of the reflection delays,  $\hat{\mathcal{S}}^{(d)}$ , are determined by identifying the indices of the  $K$  largest elements in the posterior mean of  $\mathbf{x}$ , denoted by  $\hat{\mathbf{x}}^{(d)}$ . It is obtained by exchanging  $\zeta_l^{(d-1)}$  with  $\mathbf{x}_l^{(d-1)}$  in (40). Besides, we also estimate the amplitudes of the significant components in  $\mathbf{x}$ . They can be useful to assess the sparsity level of the solution

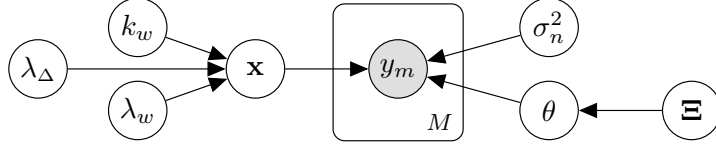


Figure 3: Dependency relations for the complete hierarchical model. The variables  $\theta$  and  $\Xi$  appear exclusively in **S2**.

and to determine the amount of optical power reflected from the FBGs. Since the posterior of  $\mathbf{x}$  is multimodal with one narrow peak around zero and another peak at some larger amplitude, the MAP is more suitable for this task. It is given by

$$\{\hat{\mathbf{x}}, \hat{\mathcal{C}}\}_{\text{MAP}}^{(d)} = \arg \max_{\mathbf{x}, \mathcal{C}} \log p(\mathbf{x}, \mathcal{C} | \mathbf{y}, \hat{\theta}^{(d-1)}) \quad (41)$$

$$\approx \arg \max_{\{\mathbf{x}_j, \mathcal{C}_j\} \in \{\mathbf{x}_l, \mathcal{C}_l\}_{l=1, \dots, L_{\text{MC}}}^{(d)}} \log p(\{\mathbf{x}_j, \mathcal{C}_j\}^{(d)} | \mathbf{y}, \hat{\theta}^{(d-1)}). \quad (42)$$

However, the estimates of  $\mathcal{S}$  obtained from  $\hat{\mathbf{x}}_{\text{MAP}}^{(d)}$  are less accurate than those obtained by the posterior mean. Therefore, the empirical MAP solution is only used to estimate the reflection amplitudes. Next, we calculate the current estimate  $\hat{\theta}^{(d)}$  by taking the expected value over  $\mathbf{x}, \mathcal{C}$  given  $\mathbf{y}, \theta$  (E-step):

$$\begin{aligned} & \mathbb{E}_{\mathbf{x}, \mathcal{C} | \mathbf{y}, \theta} \log p(\mathbf{y}, \mathbf{x}, \mathcal{C} | \theta) \\ &= \int_{\mathbb{R}_+^N} \int_{\Psi} \log p(\mathbf{y}, \mathbf{x}, \mathcal{C} | \theta) p(\mathbf{x}, \mathcal{C} | \mathbf{y}, \theta) d\mathcal{C} d\mathbf{x} \end{aligned} \quad (43)$$

$$\approx \frac{1}{L_{\text{MC}}} \sum_{l=1}^{L_{\text{MC}}} \log p(\mathbf{y}, \{\mathbf{x}_l, \mathcal{C}_l\}^{(d-1)} | \theta) \triangleq Q(\theta | \hat{\theta}^{(d-1)}). \quad (44)$$

Herein,  $\Psi$  is the product space formed by the individual domains of all variables in  $\mathcal{C}$ . In the  $M$ -step, a locally optimal value,  $\hat{\theta}^{(d)}$ , is obtained by maximizing  $\theta$  over the set  $\Theta$ , i.e.

$$\hat{\theta}^{(d)} = \arg \max_{\theta \in \Theta} Q(\theta | \hat{\theta}^{(d-1)}). \quad (45)$$

### 7.1.1 Initialization of $\theta$ via bisectional search

An adequate initialization,  $\theta^{(0)}$ , can alleviate the problem of local optima in the EM algorithm. In CFS, the desired sparsity level is known to be the number of reflections,  $K$ . Hence, a good choice for  $\theta^{(0)}$  yields a solution for  $\mathbf{x}$  with  $K$  significant non-zero elements. Starting at an arbitrary value  $\theta^{(0)} \in \Theta$ , a bisectional search within  $\Theta$  can quickly determine a suitable initial value. After choosing the first value at random,  $\Theta$  is subdivided into two parts, containing all larger and all smaller values, respectively. When the number of peaks is too high, the next trial is chosen as the median of the lower division. If it is too low, the next trial is the median of the upper division, and so on. For a properly selected  $\theta^{(0)}$ , **S1** converges faster and is more likely to approach (or even attain) the global optimum.

## 7.2 Bayesian DL: joint estimation of $(\mathbf{x}, \mathcal{C}, \theta)$ (**S2**)

In strategy **S2**, we treat  $\theta$  as a random variable. Due to its discrete nature, each element  $\theta_r \in \Theta$  is assigned a probability mass,  $p_r = p(\theta_r)$ ,  $r = 1, \dots, R_{\Theta}$ , where  $\sum_{r=1}^{R_{\Theta}} p(\theta_r) = 1$ . Then,  $\theta$  is categorically (Cat) distributed over the set of discrete dictionary parameters,  $\Theta$ , with corresponding probability masses in  $\Xi = \{p_1, \dots, p_{R_{\Theta}}\}$ . Uncertainty in the *a priori* assigned probability masses

is taken into account in terms of a prior on  $\Xi$ . The Dirichlet (Dir) distribution can be used as the conjugate prior with parameters  $\boldsymbol{\nu} = [\nu_1, \dots, \nu_{R_\Theta}]^\top$ , i.e.

$$p(\Xi) = \frac{1}{B(\boldsymbol{\nu})} \prod_{r=1}^{R_\Theta} p_r^{\nu_r}, \quad (46)$$

where  $B(\boldsymbol{\nu})$  denotes the *Beta* function and the variables  $\nu_r$ ,  $r = 1, \dots, R_\Theta$ , describe the number of occurrences of the values in  $\Theta$ . When a new element,  $\theta_q \in \Theta$ , is sampled, a posterior count is assigned to that value. After sampling another value in the next iteration, this count is reassigned to the new value. Let  $\check{\mathbf{c}} \in \mathbb{N}^{R_\Theta}$  indicate the current sample, i.e.  $c_q = 1$  for one index  $q \in \{1, \dots, R_\Theta\}$ , while all other elements are zero. A non-informative prior is obtained if all values  $\theta_q \in \Theta$  are equally likely and each element is assigned a single count. Then,  $\nu_r = 1 \forall r = 1, \dots, R_\Theta$  and a new sample has a strong impact on the posterior distribution. In contrast, for large values, e.g.  $\nu_r = 1000 \forall r = 1, \dots, R_\Theta$ , a new count leaves the distribution almost invariant. The complete model is then given by (30) and, in addition,

$$\Xi \sim \text{Dir}(\Xi | \boldsymbol{\nu}) \quad (47)$$

$$\theta | \Xi \sim \text{Cat}(\theta | R_\Theta, \Xi). \quad (48)$$

To accomplish approximate inference in this model, the variables  $\theta$  and  $\Xi$  are included in the Gibbs sampling procedure of Section 6. Therefore, the conditional distributions must be determined. Based on the dependencies in Fig. 3, and since  $\mathbf{x}$ ,  $\sigma_n^2$  and  $\theta$  are assumed to be mutually independent, we find

$$\Xi | \theta = \tilde{\Xi} \sim \text{Dir}(\Xi | \boldsymbol{\nu} + \check{\mathbf{c}}), \quad (49)$$

$$\theta | \mathbf{y}, \tilde{\Xi} \sim \text{Cat}(\theta | R_\Theta, \tilde{\Xi}). \quad (50)$$

## 8 Simulations and experimental data

Let us now evaluate the proposed sparse model and DL strategies. First, we show the qualitative behavior of the algorithms, followed by a quantitative performance analysis in comparison to the method in [62]. To this end, we consider several scenarios of different SNRs, CS sampling matrices and sample sizes. Finally, we apply our algorithms to experimental data taken from a real fiber-optic sensor.

### 8.1 Simulation setup

We consider  $K = 3$  uniform FBGs in the sensing fiber, where the observed reflections have a common amplitude,  $A_x$ , and two reflections are closely spaced. Their delays are indicated by the indices of the  $K$  most significant elements in  $\mathbf{x}$ , contained in the set  $\mathcal{S}$ . Subsequently, the dictionary parameter is re-defined relative to its true value, i.e.  $\hat{\theta}$  is replaced by  $\hat{\theta}/\theta$ . Further, we use  $R_\Theta = 100$  discrete parameter values, equally spaced between 30% and 150% of the true value. The original signal (prior to CS) contains  $L = 134$  samples of the measured photocurrent. The dictionary atoms are created using  $L$  samples of  $r(t - i\delta t)$ ,  $i = 1, \dots, N$ , with a delay spacing of  $\delta t = 50$  ns. We use two types of CS matrices,  $\Phi$ , with i.i.d. entries drawn from the distributions below:

- (a) Gauss:  $\mathcal{N}(0, 1)$ ,
- (b) DF [1]  $\{-1, 0, 1\}$  with probabilities  $\{\frac{1}{6}, \frac{2}{3}, \frac{1}{6}\}$ .

The variables  $\{\mathcal{C}, \theta, \Xi\}$  are sampled according to Section 6. For  $\mathbf{x}$  we use the 'No-U-Turn' variant of the HMC algorithm [33], which is efficiently implemented in the software package *Stan* [55]. The algorithm **S1** is initialized based on a bisectional search and runs at most  $d_{\max} = 35$  iterations. In **S2**, we use a non-informative prior for  $\theta$ , with  $p(\theta_r) = 1/R_\Theta$  and  $\nu_r = 1 \forall r = 1, \dots, R_\Theta$ .

---

**Algorithm:** Sparse estimation and PDL, strategy **S1** & **S2**


---

**Input:**  $\mathbf{y}, M, \Phi, N, L, T_d, \delta t, r(t, \theta), K, L_{\text{MC}}, d_{\text{max}}$   
**Output:**  $\hat{\mathcal{S}}, \hat{\mathbf{x}}, \hat{\theta}, \hat{\sigma}_n, d, ee$   
**Parameters:**  $a, a', a'', a_\sigma, b, b', b'', b_\sigma, d', \nu, R_\Theta, \{\theta_r\}_{r=1}^{R_\Theta}$ ,  
 internal HMC parameters (c.f. [33, 47]).  
**0. Initialize:**  $\theta$  at random  $\rightarrow \hat{\theta}^{(0)}$  via bisectional search,  
 $\mathbf{A}(\hat{\theta}^{(0)}), \{\hat{\mathbf{x}}^{(0)}, \hat{\mathcal{C}}^{(0)}\}$  as in (30), (**S2**):  $d_{\text{max}}=1$   
**1. for**  $d = 1$  to  $d_{\text{max}}$  **do**  
**2.   for**  $l = 1$  to  $L_{\text{MC}}$  **do**  
**3.     Gibbs sampling:** (i)  $\mathcal{C}_l^{(d)}$  using (33) and (34),  
                              (ii)  $\mathbf{x}_l^{(d)}$  via HMC.  
                              (**S2**): (iii)  $\theta_l^{(d)}, \Xi_l^{(d)}$  using (49) and (50)  
**4.   end for**  
**5.   Estimate:**  $\hat{\mathcal{S}}^{(d)}$  from  $\hat{\mathbf{x}}^{(d)}$  in (40) with  $\zeta_l^{(d)} \rightarrow \mathbf{x}_l^{(d)}$ ,  
                   $\hat{\mathcal{C}}^{(d)}$  from (40),  $\hat{\mathbf{x}}_{\text{MAP}}^{(d)}$  from (42),  
**5.a       (**S1**):**  $\hat{\theta}^{(d)} = \arg \max_{\theta \in \Theta} Q(\theta | \hat{\theta}^{(d-1)})$ .  
**5.b       (**S2**):**  $\hat{\theta}^{(d)}$  from (40) with  $\zeta_l^{(d)} \rightarrow \theta_l^{(d)}$ .  
**6.   if**  $\hat{\theta}^{(d)} == \hat{\theta}^{(d-1)}$  **or**  $d == d_{\text{max}}$   
**7.     return**  $\hat{\mathcal{S}}^{(d)}, \hat{\mathbf{x}}_{\text{MAP}}^{(d)}, \hat{\mathcal{C}}^{(d)}, \hat{\theta}^{(d)}, ee = \|\mathbf{y} - \Phi \mathbf{A}(\hat{\theta}^{(d)})\|_2^2$ .  
**8.   end if**  
**9. end for**

---

## 8.2 Visualization and Working Principle

The working principle of the algorithms is presented for SNR = 20 dB and a Gaussian CS matrix using  $M/L = 50\%$  of the original samples. Fig. 4 (top left) depicts the MAP solution for  $\mathbf{x}$ , obtained by HMC within Gibbs sampling according to Section 6, where  $\theta$  is fixed to the true value. It shows, that collective shrinkage, imposed by the local similarity assumption in the joint prior density of  $\mathbf{x}$ , yields a highly improved sparsity level in the presence of strong dictionary coherence. Fig. 4 (top right) shows the posterior density of  $\mathbf{x}$  in one dimension. For a non-significant component, it is strongly peaked around zero, and for a significant component, is multimodal with a strong mode around the true amplitude and a smaller mode around zero. The second row in Fig. 4 delineates the evolution of the EM algorithm in **S1** over several iterations. Fig. 4 (center left) shows the current MAP solutions for  $\mathbf{x}$ , i.e.  $\mathbf{x}_{\text{MAP}}^{(d)}$ , zoomed on the two left-sided peaks. Due to a bad initial value for  $\theta$ , more than  $K$  peaks appear in the first iterations. However, as the algorithm proceeds, significant peaks are formed only at the positions of the true significant components (black bullets). Fig. 4 (center right) shows, that also  $\theta$  approaches the true value. Fig. 4 (bottom left) delineates a typical shape of the function  $Q(\theta | \hat{\theta}^{(d-1)})$  of **S1** in (44), for a properly and badly chosen initial value,  $\theta^{(0)}$ . A good choice leads to faster convergence, while for a bad choice, the algorithm might get either stuck at a local optimum or requires many EM iterations before the maximum of the Q-function appears close the true value of  $\theta$ . Finally, Fig. 4 (bottom right) depicts for **S2**, the non-informative prior of  $\theta$  and a typical posterior density when  $\nu_r = 1 \forall r = 1, \dots, R_\Theta$ .

## 8.3 Performance evaluation

The performance is evaluated in terms of the root mean-squared error (RMSE). For a vector  $\mathbf{v}$  and an estimator  $\hat{\mathbf{v}}$ , it is given by  $\text{RMSE}(\mathbf{v}, \hat{\mathbf{v}}) = (\mathbb{E} \|\mathbf{v} - \hat{\mathbf{v}}\|_2^2)^{1/2}$ . We define  $\widehat{\text{RMSE}}$  as the approximation, where the expectation is replaced by averaging estimates over 100 Monte Carlo trials. We compare **S1**, **S2** to the PDL-OIAI algorithm in [62], which considers a deterministic sparse model and incorporates

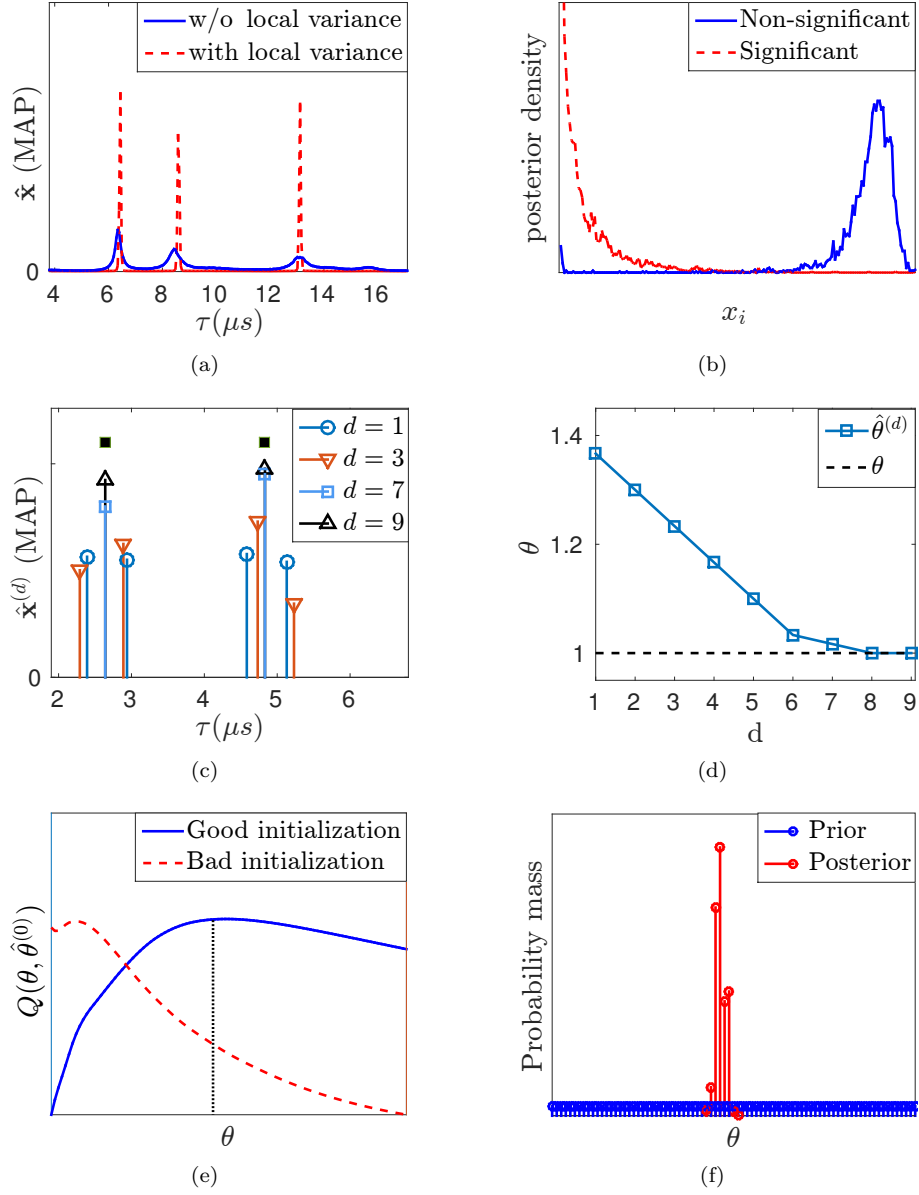


Figure 4: Visualization. Top left: empirical MAP solution for  $\mathbf{x}$  and impact of collective shrinkage. Top right: empirical posterior density of a non-/significant entry in  $\mathbf{x}$ . Center left: evolution of the temporal MAP solution for  $\mathbf{x}$  obtained by **S1**. Center right: temporal solution for  $\theta$  after the M-step. Bottom left: shape of the  $Q$ -function in **S1** for a good and bad initial value  $\theta^{(0)}$ . Bottom right: for **S2**, prior and posterior distribution of  $\theta$  for  $p(\theta_r)=1/R_\Theta$  and  $\nu_r=1, r=1, \dots, R_\Theta$ .

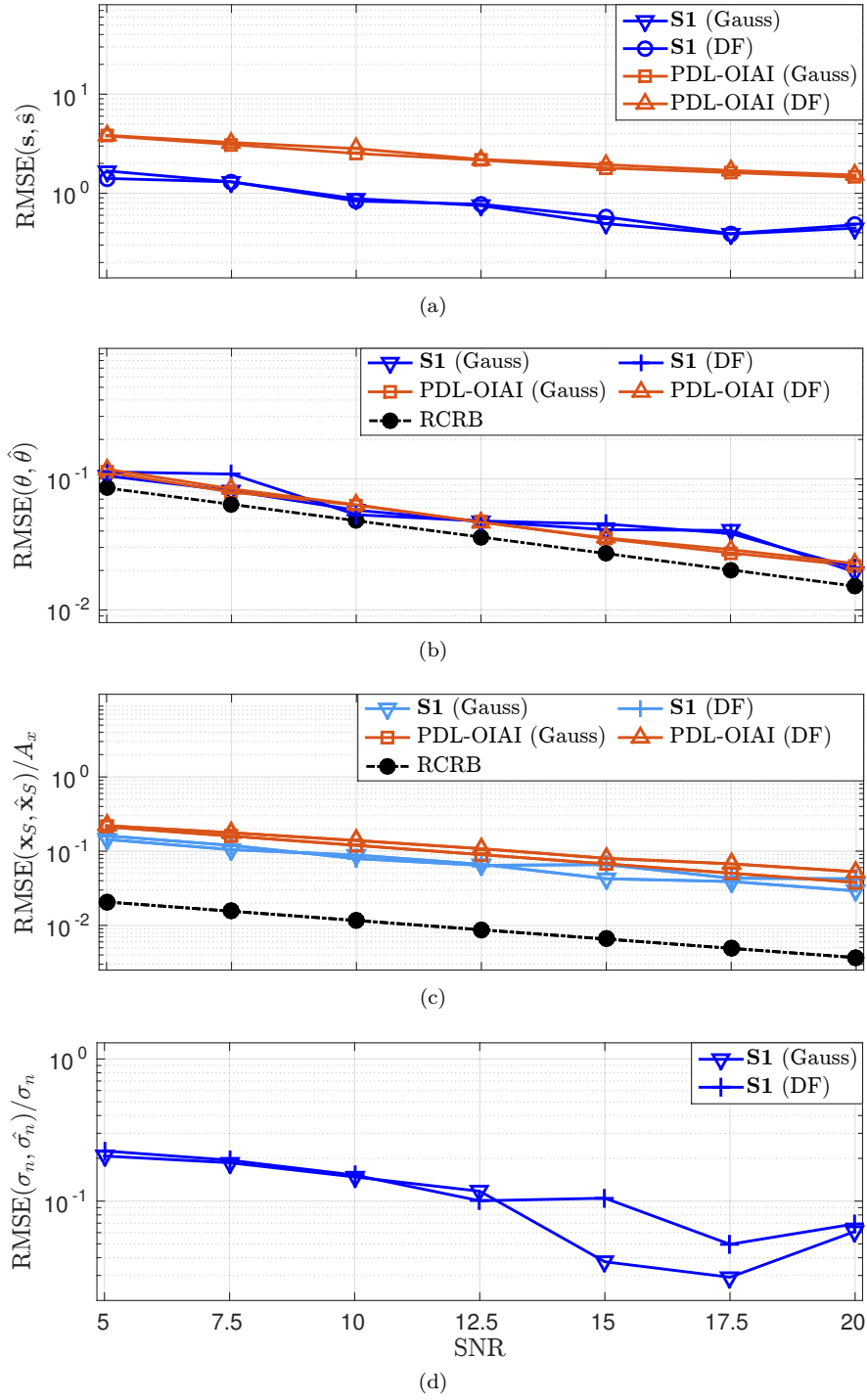


Figure 5: Performance of  $\mathbf{S1}$  in terms of the RMSE using  $M/L = 50\%$  of the original samples in comparison with PDL-OIAI in [62] and with the lower bound of the RMSE imposed by the CRB (RCRB).



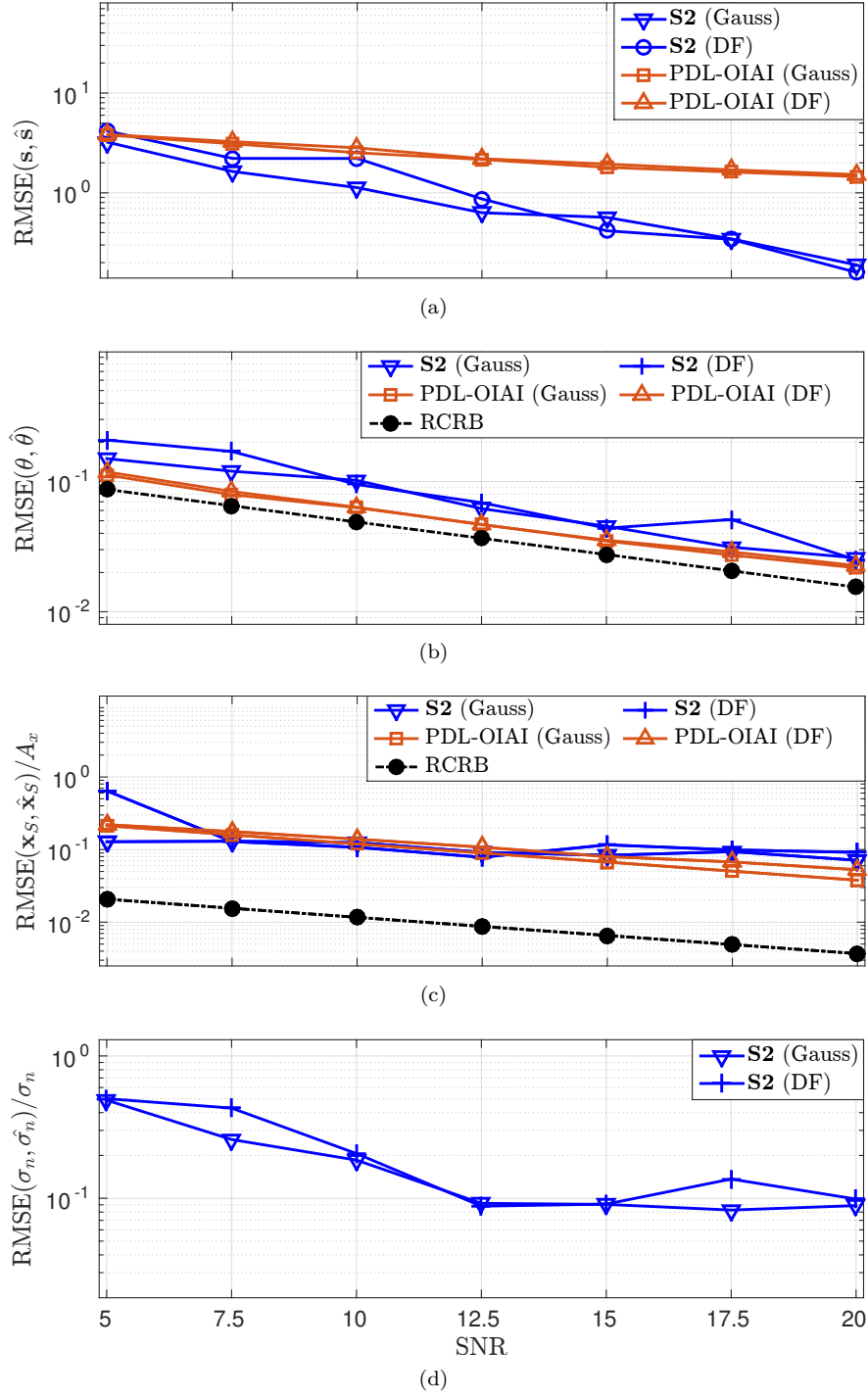


Figure 6: Performance of  $\mathbf{S2}$  in terms of the RMSE using  $M/L = 50\%$  of the original samples in comparison with PDL-OIAI in [62] and with the lower bound of the RMSE imposed by the CRB (RCRB).

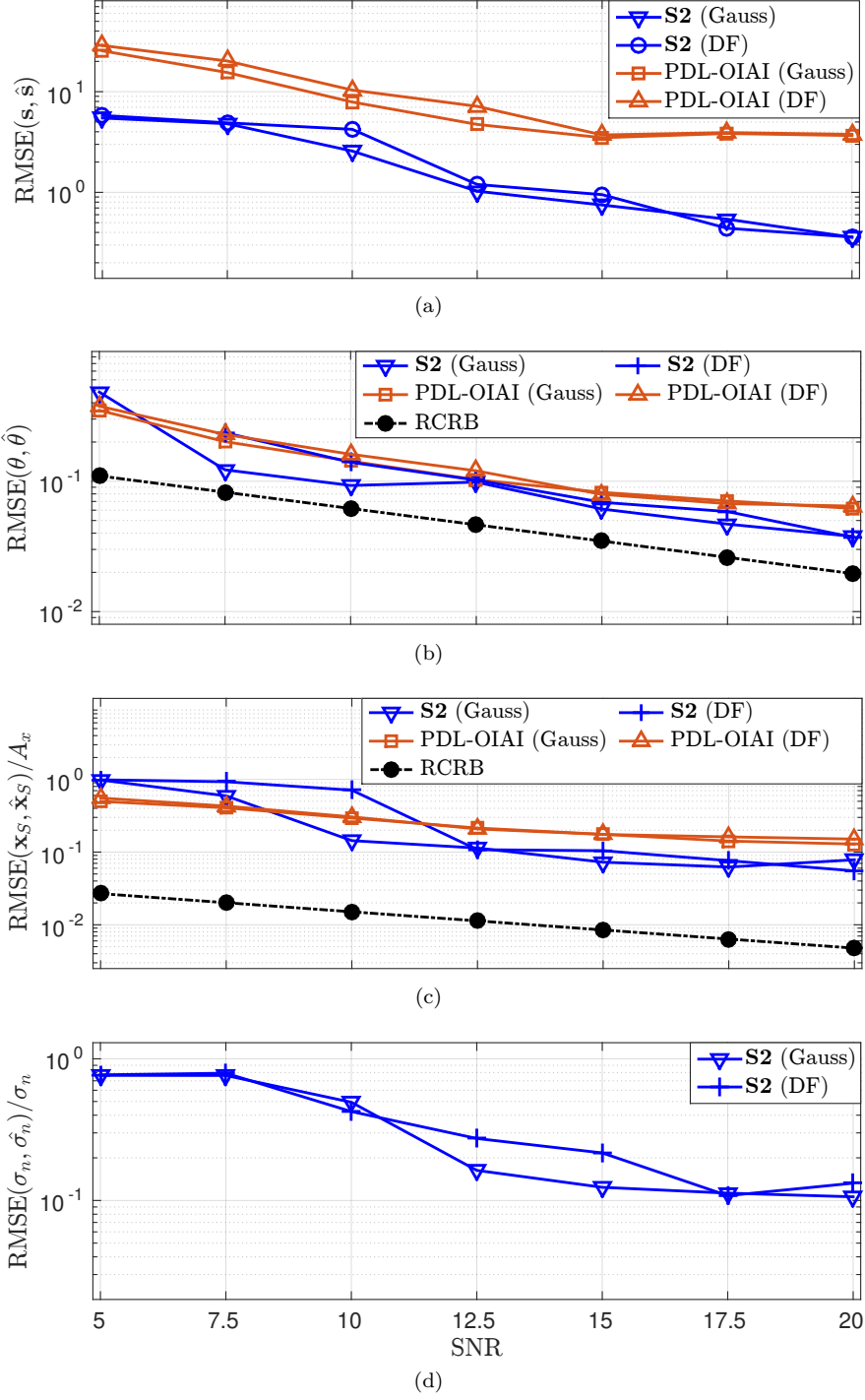


Figure 7: Performance of  $\mathbf{S2}$  in terms of the RMSE using  $M/L = 30\%$  of the original samples in comparison with PDL-OIAI in [62] and with the lower bound of the RMSE imposed by the CRB (RCRB).

a pre-processing routine to handle strong dictionary coherence. To calculate the CRB of Section 4, the derivative of  $r(t, \theta)$  with respect to  $\theta$  must be determined for all dictionary elements. Since  $r(t, \theta)$  is not a simple function of  $\theta$ , it can be approximated for a certain value  $\theta_0$ . For the  $(l, i)$ -th element in  $\mathbf{A}'(\theta)$ , we obtain

$$\left. \frac{\partial}{\partial \theta} [\mathbf{a}_i(\theta)]_l \right|_{\theta_0} \approx \frac{r(lT_d - \tau_i, \theta_0) - r(lT_d - \tau_i, \theta_0 - \Delta\theta)}{\Delta\theta}. \quad (51)$$

Fig. 5-7 show the results of the proposed and the competitive method. Herein,  $\mathbf{s} \in \mathbb{N}^K$  contains all elements in  $\mathcal{S}$ ,  $\mathbf{x}_\mathcal{S} \in \mathbb{R}_+^K$  contains the coefficients of  $\mathbf{x}$  with indices in  $\mathcal{S}$ . The  $\text{RMSE}(\mathbf{x}_\mathcal{S}, \hat{\mathbf{x}}_\mathcal{S})$  compares the estimated amplitudes at the positions in  $\hat{\mathcal{S}}$  to the true common amplitude,  $A_x$ , at positions in  $\mathcal{S}$ . The lower bound of the RMSE for jointly estimating *deterministic* parameters  $(\mathbf{x}_\mathcal{S}, \theta)$ , induced by the CRB derived in Section 4, is denoted by 'RCRB'.

Fig. 5 shows the results for  $\{\mathbf{s}, \mathbf{x}_\mathcal{S}, \theta, \sigma_n\}$ , obtained by **S1** using 50% of the original samples. For fewer samples, the EM algorithm in **S1** becomes unstable. Fig. 6-7 depicts the results obtained by **S2** using 50% and 30% of the original samples, respectively. It shows, that **S2** is more robust against small sample sizes and missing data than **S1**. Generally, the error is only marginally affected by the type of the CS sampling matrix, i.e. (a) or (b). In all scenarios, **S1** and **S2** achieve a significantly lower error in estimating  $\mathbf{s}$  than PDL-OIAI. At low SNRs, **S1** performs better than **S2**, while **S2** becomes better at high SNRs. However, PDL-OIAI estimates  $\theta$  with slightly higher accuracy than **S1** and **S2**. When 50% of the original samples are used, the error closely adheres to the RCRB. The amplitudes,  $\mathbf{x}_\mathcal{S}$ , are estimated with similar accuracy by both, **S1** and **S2**, and no improvement is achieved compared to PDL-OIAI. Also, the distance to the RCRB is almost constant at all SNRs. Regarding the noise level,  $\sigma_n^2$ , **S1** yields a slightly smaller estimation error than **S2**. PDL-OIAI does not provide a simple means for estimating  $\sigma_n^2$ , which is an advantage of **S1** and **S2**. In the presented results for PDL-OIAI, it is assumed that pure noise samples are available to estimate  $\sigma_n^2$ . The instability of the RMSE between SNRs of 15 and 17.5 dB in Fig. 5 might arise from averaging over an insufficient number of samples. It is also possible that the MCMC algorithm took longer to converge to the stationary distribution for SNR=20 dB, e.g. due to an unlucky initialization, thus, increasing the error.

## 8.4 Experimental Data

To complete our study, we apply **S1** and **S2** to experimental data taken from the real fiber sensor system in [46, 63]. It was acquired at the Yamashita laboratory of photonic communication devices at The University of Tokyo, Japan. We consider  $L = 134$  original samples of the received sensor signal and use  $M/L = 50\%$  of the original samples. The delay spacing between the  $N=2L$  dictionary atoms is  $\delta t \approx 50$  ns. The sensing fiber contains  $K = 4$  FBGs and the delays of the reflected signals are potentially off-grid. Their positions are approximately at  $[7.79, 9.05, 10.27, 12.30] \mu\text{s}$ . We perform 100 Monte Carlo trials to estimate  $\{\mathcal{S}, \mathbf{x}_\mathcal{S}, \theta\}$ . Fig. 8 (left) shows the the original sensor signal and one estimated reflection from FBG<sub>3</sub>. The shaded area indicates the standard deviation in estimating  $\theta$ . **S1** estimates a narrower reflection, which also results in slightly different estimates of  $\mathcal{S}$ . Fig. 8 (right) depicts  $\hat{\mathbf{x}}_\mathcal{S}$  at the estimated positions in  $\hat{\mathcal{S}}$ . The shaded areas represents the standard deviation for  $\mathcal{S}$  and the vertical error bars indicate the standard deviation for  $\mathbf{x}_\mathcal{S}$ . Essentially, the results of **S1** and **S2** are comparable, although the variance in the estimates of  $\mathcal{S}$  is marginally smaller in the case of **S2**. Similar performance was reported for PDL-OIAI in [62].

## 9 Discussion

Based on simulations and experimental results, we demonstrate that the proposed sparsity model and our DL strategies, **S1**, **S2**, are useful in CFS and can be used for an automated estimation of the reflection delays. In comparison to PDL-OIAI in [62], where the underlying model treats  $\mathbf{x}$  and  $\theta$  as deterministic parameters, the following general observation can be made: The methods **S1** and **S2**, based on a probabilistic sparse model, show comparable performance to PDL-OIAI but do not exceed the performance limit imposed by the non-Bayesian CRB. However, a significant improvement

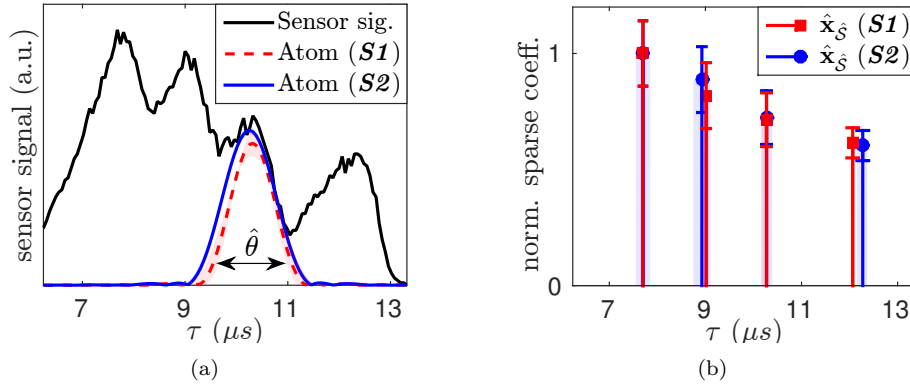


Figure 8: Real data example for  $\mathbf{S1}$  and  $\mathbf{S2}$ : Sensor signal with estimated reflections, where the shaded regions indicate the standard deviation for estimating  $\theta$  (left). Estimated sparse signal, where the shaded areas show the standard error in estimating the reflection delays, indicated by  $\mathcal{S}$ , and the vertical error bars show the standard error for estimating  $\mathbf{x}_{\mathcal{S}}$  (right).

is achieved in estimating  $\mathcal{S}$ . It should be emphasized that  $\mathcal{S}$  is of major importance in WDM-based CFS. It indicates the reflection delays, that are used to infer the quantity or nature of impairments at the FBGs. The amplitudes,  $\mathbf{x}_{\mathcal{S}}$ , can be used to determine the sparsity level and the amount of optical power reflected from the FBGs. We find that all competing methods estimate  $\mathbf{x}_{\mathcal{S}}$  similarly accurate. The real data example shows that  $\mathbf{S1}$ ,  $\mathbf{S2}$  are insensitive to signal features that are not explicitly modeled, e.g. the skewness of the reflections or a signal-dependent noise amplitude. This was also reported for PDL-OIAI [62].

Our results for  $\mathcal{S}$  indicate that the proposed sparsity model is better able to handle strong dictionary coherence than PDL-OIAI, which adopts a dictionary pre-processing routine to reduce the dictionary coherence. We ascribe this ability in part to the favorable selective shrinkage properties of the Weibull prior. Such behavior was previously reported for general heavy-tailed priors in [50, 54]. Regarding the relation to non-convex optimization, we find that constraints imposed on the  $\ell_p$ -norm, with  $0 < p < 1$ , are indeed useful in the presence of strong dictionary coherence. In this context, we support the findings in [17, 18], that report relaxed RIP conditions when  $\ell_1$ -minimization is replaced by non-convex optimization methods. Another important factor, that contributes to the ability of handling strong dictionary coherence, is the local similarity model introduced in the joint prior density of  $\mathbf{x}$ . We observe much sparser solutions due to its collective shrinkage property, without the need for any dictionary pre-processing as in PDL-OIAI. Although this model is designed to deal with the unique features of the CFS dictionary, it can be used for general shift-invariant dictionaries with similar structures and high coherence levels. Therefore, it offers a broader applicability beyond the CFS problem. For parametric DL, all compared methods seem equally suitable for estimating  $\theta$ , but  $\mathbf{S2}$  and PDL-OIAI are more stable for small sample sizes. Since the type of the CS matrix has only marginal impact, DF matrices in [1] are favorable. They are easy to implement, require low storage, and reduce the average sampling rate by 66%, since 2/3 of all projections are zero.

The computational complexity of  $\mathbf{S1}$  and  $\mathbf{S2}$  is dominated by drawing samples from the posterior of  $\mathbf{x}$  using HMC. HMC shows high efficacy in sampling this high-dimensional space in the presence of correlation. It yields samples from the desired posterior, that are weakly sparse with sharp peaks only close to the true positions of the significant components. Compared to optimization methods such as PDL-OIAI, MCMC is slower (c.f. [43]) but some preliminary efforts are necessary for choosing a proper regularization parameter in the  $\ell_1$ -minimization problem. The run-time complexity of PDL-OIAI is dominated by a costly but essential data-dependent pre-processing routine to deal with severe dictionary coherence. This can be implemented using parallel processing and might be more efficient in situations, where CFS is used for permanent perturbation monitoring. Nonetheless, it requires an initial estimate of the non-perturbed reference system. For this task,  $\mathcal{S}$  can be more accurately es-

timated using **S1** and **S2**. In contrast to PDL-OIAI, they are also able to estimate the noise level. Combining these methods for calibration and permanent monitoring is a promising perspective for practical systems. A limiting factor in **S1** and **S2** is the MCMC runtime, i.e. the number of available samples for Monte Carlo integration. Depending on the initial point, sufficient time has to be given for the algorithms to converge to the stationary distribution. Also, **S1** may get stuck in local optima but the proposed initialization using a bisectional search can lower this chance and helps to speed up the convergence of the algorithm. A possible extension of this work can include multiple CS sample vectors to improve the SNR conditions. This might yield more accurate results and stable behavior. A similar technique was proposed in [41]. Also, as pointed out in [62], additional local dictionary parameters can be considered. Since the reflections in the experimental data are non-uniform, this might improve both robustness and accuracy.

## 10 Conclusion

We present a sparse estimation and parametric dictionary learning framework for Compressed Fiber Sensing (CFS) based on a probabilistic hierarchical sparse model. The significant components in the sparse signal indicate reflection delays, that can be used to infer the quantity and nature of external impairments. In order to handle severe dictionary coherence and to accommodate specific characteristics of the signal, a Weibull prior is employed to promote selective shrinkage. This choice can be related to non-convex optimization based on the  $\ell_p$ -norm. To further alleviate the problem of dictionary coherence, we leverage the particular structure of the dictionary and assign a local variance to the differential sparse coefficients. This model can be useful for general shift-invariant dictionaries with similar structure and strong coherence. We propose two parametric dictionary learning strategies, **S1** and **S2**, to estimate the dictionary parameter,  $\theta$ . In **S1**,  $\theta$  is treated as a deterministic parameter and estimated using a Monte Carlo EM algorithm. In **S2**, a probabilistic hierarchical model for  $\theta$  is considered. A hybrid MCMC method based on Gibbs sampling and Hamilton Monte Carlo is used for approximate inference. In simulations and by experimental data, we show the applicability and efficacy of the proposed sparse model, together with the methods **S1** and **S2**, for an automated estimation of the reflection delays and the dictionary parameter in CFS. In a comparative analysis with an existing method, based on a deterministic sparse model, we highlight advantages, disadvantages and limitations, that can serve as a guidance to choose an adequate method for practical systems. To better assess the performance gain of a probabilistic sparse model, the Cramér-Rao bound is derived for the joint estimation of deterministic sparse coefficients and the dictionary parameter in CFS. Drawbacks of the proposed methods are the generally high computational costs of MCMC methods, and the lack of simple diagnostic tools for Markov chain convergence and sample independence. Also, **S1** suffers from the problem of local optima. As a remedy, we propose a bisectional search to find a proper initialization. In subsequent investigations, multiple CS sample vectors and additional local dictionary parameters can be taken into account. Also, variational Bayes methods can be used to speed up computations.

## Acknowledgment

This work was supported by the 'Excellence Initiative' of the German Federal and State Governments and the Graduate School of Computational Engineering at Technische Universität Darmstadt. The authors would like to thank Professor S. Yamashita and his group at The University of Tokyo, Japan, for kindly providing experimental data of the fiber sensor in [63].

## References

- [1] D. Achlioptas. Database-friendly random projections: Johnson-lindenstrauss with binary coins. *J. Comput. Syst. Sci.*, 66(4):671–687, 2003.

- [2] A. Agarwal, A. Anandkumar, P. Jain, P. Netrapalli, and R. Tandon. Learning sparsely used overcomplete dictionaries. *J. Mach. Learn. R. Workshop (JMLR)*, 35:1–15, 2014.
- [3] Y. Altmann, M. Pereyra, and J. Bioucas-Dias. Collaborative sparse regression using spatially correlated supports - application to hyperspectral unmixing. *IEEE Trans. Image Process.*, 24(12):5800–5811, 2015.
- [4] M. Ataee, H. Zayyani, M. Babaie-Zadeh, and C. Jutten. Parametric dictionary learning using steepest descent. In *IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP)*, pages 1978–1981, Mar 2010.
- [5] R. G. Baraniuk. Compressive sensing [lecture notes]. *IEEE Signal Process. Mag.*, 24(4):118–121, Jul 2007.
- [6] A. Beck and L. Tetruashvili. On the convergence of block coordinate descent type methods. *SIAM J. Optim.*, 23(4):2037–2060, 2013.
- [7] Z. Ben-Haim and Y. C. Eldar. The Cramér-Rao bound for estimating a sparse parameter vector. *IEEE Trans. Signal Process.*, 58(6):3384–3389, Jun 2010.
- [8] C. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.
- [9] T. Blumensath and M. Davies. Sparse and shift-invariant representations of music. *IEEE Trans. on Audio, Speech, Language Process.*, 14(1):50–57, Jan 2006.
- [10] B. Z. Bobrovsky, E. Mayer-Wolf, and M. Zakai. Some classes of global Cramér-Rao bounds. *Ann. Stat.*, 15(4):1421–1438, 1987.
- [11] E. J. Candes. The restricted isometry property and its implications for compressed sensing. *C.R. Math.*, 346:589–592, 2008.
- [12] E. J. Candes, Y. C. Eldar, D. Needell, and P. Randall. Compressed sensing with coherent and redundant dictionaries. *Appl. Comput. Harmon. Anal.*, 31:59–73, Jul 2011.
- [13] E. J. Candes, J. Romberg, and T. Tao. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory*, 52(2):489–509, 2006.
- [14] E. J. Candes and T. Tao. Decoding by linear programming. *IEEE Trans. Inf. Theory*, 51(12):4203–4215, Dec 2005.
- [15] E. J. Candes and M. B. Wakin. An introduction to compressive sampling. *IEEE Signal Process. Mag.*, 25(2):21–30, Mar 2008.
- [16] C. M. Carvalho, N. G. Polson, and J. G. Scott. The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480, 2010.
- [17] R. Chartrand. Exact reconstruction of sparse signals via nonconvex minimization. *IEEE Signal Process. Lett.*, 14(10):707–710, Oct 2007.
- [18] R. Chartrand and V. Staneva. Restricted isometry properties and nonconvex compressive sensing. *Inverse Problems*, 24(3):035020, 2008.
- [19] C. Chen, D. Carlson, Z. Gan, C. Li, and L. Carin. Bridging the gap between stochastic gradient MCMC and stochastic optimization. In *Proc. 19th Int. Conf. Artificial Intell. Stat.*, pages 1051–1060, 2016.
- [20] C. Chen, N. Ding, and L. Carin. On the convergence of stochastic gradient MCMC algorithms with high-order integrators. In *Adv. Neural Inf. Process Syst. 28*, pages 2278–2286. Curran Associates, Inc., 2015.

- [21] J. Chen, F. D. Bushman, J. D. Lewis, G. D. Wu, and H. Li. Structure-constrained sparse canonical correlation analysis with an application to microbiome data analysis. *Biostatistics*, 14:244–258, 2013.
- [22] T. Chen, E. Fox, and C. Guestrin. Stochastic gradient Hamiltonian Monte Carlo. In *Proc. 31st Int. Conf. Machine Learn.*, pages 1683–1691, 2014.
- [23] B. Culshaw and A. Kersey. Fiber-optic sensing: A historical perspective. *J. Lightwave Technol.*, 26(9):1064–1078, May 2008.
- [24] D. L. Donoho and M. Elad. Optimally sparse representation in general (nonorthogonal) dictionaries via  $\ell_1$ -minimization. *Proc. Nat. Acad. Sci. (PNAS)*, 100(5):2197–2202, 2003.
- [25] D.L. Donoho and X. Huo. Uncertainty principles and ideal atomic decomposition. *IEEE Trans. Inf. Theory*, 47(7):2845–2862, Nov 2001.
- [26] Y. C. Eldar, P. Kuppinger, and H. Bolcskei. Block-sparse signals: Uncertainty relations and efficient recovery. *IEEE Trans. Signal Process.*, 58(6):3042–3054, Jun 2010.
- [27] Y. C. Eldar and G. Kutyniok, editors. *Compressed Sensing: Theory and Applications*. Cambridge University Press, 2012.
- [28] D. Fink. A compendium of conjugate priors, 1997.
- [29] K. Fyhn, M. F. Duarte, and S. H. Jensen. Compressive parameter estimation for sparse translation-invariant signals using polar interpolation. *IEEE Trans. Signal Process.*, 63(4):870–881, Feb 2015.
- [30] M. D. Gupta and S. Kumar. Non-convex p-norm projection for robust sparsity. In *IEEE Int. Conf. Comput. Vision*, pages 1593–1600, Sydney, Australia, 2013.
- [31] T. L. Hansen, M. A. Badiu, B. H. Fleury, and B. D. Rao. A sparse Bayesian learning algorithm with dictionary parameter estimation. In *IEEE 8th Sensor Array and Multichannel Signal Process. Workshop (SAM)*, pages 385–388, June 2014.
- [32] N. J. Higham. *Accuracy and Stability of Numerical Algorithms*. SIAM, 2 edition, 2002.
- [33] M. D. Homan and A. Gelman. The no-u-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *J. Mach. Learn. Res.*, 15(1):1593–1623, Jan 2014.
- [34] H. Ishwaran and J. S. Rao. Spike and slab variable selection: Frequentist and bayesian strategies. *Ann. Statist.*, 33(2):730–773, 2005.
- [35] H. Ishwaran and S. Rao. Spike and slab gene selection for multigroup microarray data. *J. Amer. Statist. Assoc.*, 100:764–780, 2005.
- [36] P. Jost, P. Vanderghenst, S. Lesage, and R. Gribonval. MoTIF: An efficient algorithm for learning translation invariant dictionaries. In *IEEE Int. Conf. on Acoust. Speech, Signal Process.*, volume 5, pages V–V, May 2006.
- [37] S. M. Kay. *Fundamentals of Statistical Signal Processing, Volume I: Estimation Theory*. Prentice Hall, 1993.
- [38] A. D. Kersey, M. A. Davis, H. J. Patrick, M. LeBlanc, K. P. Koo, C. G. Askins, M. A. Putnam, and E. J. Friebele. Fiber grating sensors. *J. Lightwave Technol.*, 15(8):1442–1463, Aug 1997.
- [39] M. Leigsnering, F. Ahmad, M. G. Amin, and A. M. Zoubir. Parametric dictionary learning for sparsity-based TWRI in multipath environments. *IEEE Trans. Aerosp. Electron. Syst.*, 52(2):532–547, Apr 2016.

- [40] J. Mairal, M. Elad, and G. Sapiro. Sparse representation for color image restoration. *IEEE Trans. Image Proc.*, 17(1):53–69, 2008.
- [41] D. Malioutov, M. Cetin, and A. S. Willsky. A sparse signal reconstruction perspective for source localization with sensor arrays. *IEEE Trans. Signal Process.*, 53(8):3010–3022, 2005.
- [42] R. M. Measures, M. LeBlanc, K. Liu, S. Ferguson, T. Valis, D. Hogg, R. Turner, and K. McEwen. Fiber optic sensors for smart structures. *Opt. Lasers Eng.*, 16:127–152, 1992. Special Issue on Optical Sensors for Aerospace Applications.
- [43] S. Mohamed, K. A. Heller, and Z. Ghahramani. Bayesian and  $l_1$  approaches for sparse unsupervised learning. In John Langford and Joelle Pineau, editors, *Proc. 29th Int. Conf. Mach. Learn. (ICML-12)*, pages 751–758, New York, USA, 2012. ACM.
- [44] A. Mohammad-Djafari. Bayesian approach with prior models which enforce sparsity in signal and image processing. *EURASIP J. Adv. Signal Process.*, 2012(1):1–19, 2012.
- [45] K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT Press, 2012.
- [46] Y. Nakazaki and S. Yamashita. Fast and wide tuning range wavelength-swept fiber laser based on dispersion tuning and its application to dynamic FBG sensing. *Opt. Express*, 17(10):8310–8318, May 2009.
- [47] R. M. Neal. *MCMC using Hamilton Dynamics*. Handbook of Markov Chain Monte Carlo. Chapman and Hall / CRC, 2011. Chapter 5, edited by S. Brooks, A. Gelman, G. Jones and X.-L. Meng.
- [48] O’Hagan. Bayesian inference. In *Kendall’s Advanced Theory of Statistics*, volume B, 1994.
- [49] Y. C. Pati, R. Rezaifar, and P. S. Krishnaprasad. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *27th Asilomar Conf. Signals Syst. Comput.*, volume 1, pages 40–44, Nov 1993.
- [50] N. G. Polson and J. G. Scott. Shrink globally, act locally: Sparse Bayesian regularization and prediction. In *Bayesian Statistics*, volume 9, 2010.
- [51] H. Raja, W. U. Bajwa, F. Ahmad, and M. G. Amin. Parametric dictionary learning for TWRI using distributed particle swarm optimization. In *Proc. IEEE Radar Conf.*, May 2016.
- [52] H. Rauhut, K. Schnass, and P. Vandergheynst. Compressed sensing and redundant dictionaries. *IEEE Trans. Inf. Theory*, 54(5):2210–2219, May 2008.
- [53] R. Rubinstein, A. M. Bruckstein, and M. Elad. Dictionaries for sparse representation modeling. *Proc. IEEE*, 98(6):1045–1057, Jun 2010.
- [54] M. W. Seeger. Bayesian inference and optimal design for the sparse linear model. *J. Mach. Learn. Res.*, 9:759–813, Jun 2008.
- [55] Stan Development Team. Stan: A C++ library for probability and sampling, version 2.9.0, 2015.
- [56] R. Tibshirani. Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Series B*, 58:267–288, 1994.
- [57] J. A. Tropp. Greed is good: algorithmic results for sparse approximation. *IEEE Trans. Inf. Theory*, 50(10):2231–2242, Oct 2004.
- [58] E. Udd. Fiber optic smart structures. *Proc. IEEE*, 84(1):60–67, Jan 1996.
- [59] M. B. Wakin, S. Sarvotham, M. F. Duarte, D. Baron, and R. G. Baraniuk. Recovery of jointly sparse signals from few random projections. In *Proc. Workshop Neural Inform. Process. Syst. (NIPS)*, pages 1435–1442, Vancouver, Canada, Dec 2005.



- [60] C. Weiss and A. M. Zoubir. Fiber sensing using UFWT-lasers and sparse acquisition. In *Proc. 21st Eur. Signal Process. Conf. (EUSIPCO)*, pages 1–5. EURASIP, 2013.
- [61] C. Weiss and A. M. Zoubir. Fiber sensing using wavelength-swept lasers: A compressed sampling approach. In *3rd Int. Workshop Compressed Sens. Theory Appl. Radar Sonar Remote Sens. (CoSeRa)*, pages 21–25, Jun 2015.
- [62] C. Weiss and A. M. Zoubir. A compressed sampling and dictionary learning framework for WDM-based distributed fiber sensing. In *arXiv:1609.08043 [stat.ME]*, 2016.
- [63] S. Yamashita, Y. Nakazaki, R. Konishi, and O. Kusakari. Wide and fast wavelength-swept fiber laser based on dispersion tuning for dynamic sensing. *J. Sensors*, 2009:12, 2009.
- [64] A. Yuan. Bayesian frequentist hybrid inference. *Ann. Stat.*, 37(5A):2458–2501, 2009.
- [65] A. Yuan, X. Zhang, and G. Han. Review of hybrid Bayesian inference and its applications. *Ann. Biometrics Biostat.*, 2(3):1023, 2015.
- [66] Z. Zhang and B. D. Rao. Sparse signal recovery with temporally correlated source vectors using sparse bayesian learning. *IEEE J. Sel. Topics Signal Process.*, 5(5):912–926, Sep 2011.